

# GAUSSIAN APPROXIMATION OF SUPREMA OF EMPIRICAL PROCESSES\*

BY VICTOR CHERNOZHUKOV<sup>†</sup>, DENIS CHETVERIKOV<sup>‡</sup> AND KENGO KATO<sup>§</sup>

*MIT<sup>†</sup>, UCLA<sup>‡</sup>, and University of Tokyo<sup>§</sup>*

This paper develops a new direct approach to approximating suprema of general empirical processes by a sequence of suprema of Gaussian processes, without taking the route of approximating whole empirical processes in the sup-norm. We prove an abstract approximation theorem applicable to a wide variety of statistical problems, such as construction of uniform confidence bands for functions. Notably, the bound in the main approximation theorem is non-asymptotic and the theorem allow for functions that index the empirical process to be unbounded and have entropy divergent with the sample size. The proof of the approximation theorem builds on a new coupling inequality for maxima of sums of random vectors, the proof of which depends on an effective use of Stein's method for normal approximation, and some new empirical process techniques. We study applications of this approximation theorem to local and series empirical processes arising in nonparametric estimation via kernel and series methods, where the classes of functions change with the sample size and are non-Donsker. Importantly, our new technique is able to prove the Gaussian approximation for the supremum type statistics under weak regularity conditions, especially concerning the bandwidth and the number of series functions, in those examples.

**1. Introduction.** This paper is concerned with the problem of approximating suprema of empirical processes by a sequence of suprema of Gaussian processes. To formulate the problem, let  $X_1, \dots, X_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$  with common distribution  $P$ . Suppose that there is a sequence  $\mathcal{F}_n$  of classes of measurable functions  $S \rightarrow \mathbb{R}$ , and consider the empirical process indexed by  $\mathcal{F}_n$ :

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_1)]), \quad f \in \mathcal{F}_n.$$

---

\*First arXiv version: December 31, 2012. Revised April 26, 2014. V. Chernozhukov and D. Chetverikov are supported by a National Science Foundation grant. K. Kato is supported by the Grant-in-Aid for Young Scientists (B) (25780152), the Japan Society for the Promotion of Science.

*Keywords and phrases:* coupling, empirical process, Gaussian approximation, kernel estimation, local empirical process, series estimation, supremum

For a moment, we implicitly assume that each  $\mathcal{F}_n$  is “nice” enough and postpone the measurability issue. This paper tackles the problem of approximating  $Z_n = \sup_{f \in \mathcal{F}_n} \mathbb{G}_n f$  by a sequence of random variables  $\tilde{Z}_n$  equal in distribution to  $\sup_{f \in \mathcal{F}_n} B_n f$ , where each  $B_n$  is a centered Gaussian process indexed by  $\mathcal{F}_n$  with covariance function  $\mathbb{E}[B_n(f)B_n(g)] = \text{Cov}(f(X_1), g(X_1))$  for all  $f, g \in \mathcal{F}_n$ . We look for conditions under which there exists a sequence of such random variables  $\tilde{Z}_n$  with

$$(1) \quad |Z_n - \tilde{Z}_n| = O_{\mathbb{P}}(r_n),$$

where  $r_n \rightarrow 0$  as  $n \rightarrow \infty$  is a sequence of constants. These results have immediate statistical implications; see Remark 2.5 and Section 3 ahead.

The study of asymptotic and non-asymptotic behaviors of the supremum of the empirical process is one of the central issues in probability theory, and dates back to the classical work of [34]. The (tractable) distributional approximation of the supremum of the empirical process is of particular importance in mathematical statistics. A leading example is uniform inference in nonparametric estimation, such as construction of uniform confidence bands and specification testing in nonparametric density and regression estimation where critical values are given by quantiles of supremum type statistics [see, e.g., 2, 37, 52, 29, 28, 12]. Another interesting example appears in econometrics where there is an interest in estimating a parameter that is given as the extremum of an unknown function such as a conditional mean function. [14] proposed a precision-corrected estimate for such a parameter. In construction of their estimate, approximation of quantiles of a supremum type statistic is needed, to which the Gaussian approximation plays a crucial role.

A related but different problem is that of approximating *whole* empirical processes by a sequence of Gaussian processes in the sup-norm. This problem is more difficult than (1). Indeed, (1) is implied if there exists a sequence of versions of  $B_n$  (which we denote by the same symbol  $B_n$ ) such that

$$(2) \quad \|\mathbb{G}_n - B_n\|_{\mathcal{F}_n} := \sup_{f \in \mathcal{F}_n} |(\mathbb{G}_n - B_n)f| = O_{\mathbb{P}}(r_n).$$

There is a large literature on the latter problem (2). Notably, Komlós et al. [36] (henceforth, abbreviated as KMT) proved that  $\|\mathbb{G}_n - B_n\|_{\mathcal{F}} = O_{a.s.}(n^{-1/2} \log n)$  for  $S = [0, 1]$ ,  $P =$  uniform distribution on  $[0, 1]$ , and  $\mathcal{F} = \{1_{[0,t]} : t \in [0, 1]\}$ . See [42] and [5] for refinements of KMT’s result. [43], [35] and [52] developed extensions of the KMT construction to more general classes of functions.

The KMT construction is a powerful tool in addressing the problem (2), but when applied to general empirical processes, it typically requires strong

conditions on classes of functions and distributions. For example, Rio [52] required that  $\mathcal{F}_n$  are uniformly bounded classes of functions having uniformly bounded variations on  $S = [0, 1]^d$ , and  $P$  has a continuous and positive Lebesgue density on  $[0, 1]^d$ . Such conditions are essential to the KMT construction since it depends crucially on the Haar approximation and binomial coupling inequalities of Tusnády. Note that [35] directly made an assumption on the accuracy of the Haar approximation of the class of functions, but still required similar side conditions to [52] in concrete applications; see Section 11 in [35]. [20], [1] and [54] considered the problem of Gaussian approximation of general empirical processes with different approaches and thereby without such side conditions. [20] used a finite approximation of a (possibly uncountably) infinite class of functions and apply a coupling inequality of [61] to the discretized empirical process (more precisely, [20] used a version of Yurinskii's inequality proved by [18]). [1] and [54], on the other hand, used a coupling inequality of [62] instead of Yurinskii's and some recent empirical process techniques such as Talagrand's [57] concentration inequality, which leads to refinements of Dudley and Philipp's results in some cases. However, the rates that [18], [1] and [54] established do not lead to tight conditions for the Gaussian approximation in non-Donsker cases, with important examples being the suprema of empirical processes arising in nonparametric estimation, namely the suprema of local and series empirical processes (see Section 3 for detailed treatment).

We develop here a new direct approach to the problem (1), without taking the route of approximating the whole empirical process in the sup-norm and with different technical tools than those used in the aforementioned papers (especially the approach taken does not rely on the Haar expansion and hence differs from the KMT type approximation). We prove an abstract approximation theorem (Theorem 2.1) that leads to results of type (1) in several situations. The proof of the approximation theorem builds on a number of technical tools that are of interest in their own rights: notably, 1) a new coupling inequality for maxima of sums of random vectors (Theorem 4.1), where Stein's method for normal approximation (building here on [7] and originally due to [55, 56]) plays an important role (see also [51, 44, 9]); 2) a deviation inequality for suprema of empirical processes that only requires finite moments of envelope functions (Theorem 5.1), due essentially to the recent work of [4], complemented with a new "local" maximal inequality for the expectation of suprema of empirical processes that extends the work of [60] (Theorem 5.2). We study applications of this approximation theorem to local and series empirical processes arising in nonparametric estimation via kernel and series methods, and demonstrate that our new technique is able

to provide the Gaussian approximation for the supremum type statistics under weak regularity conditions, especially concerning the bandwidth and the number of series functions, in those examples. A companion work [12] provides multiplier bootstrap methods for (approximate and valid) computation of Gaussian approximations  $\tilde{Z}_n$  in applications (see also Remark 3.3 below).

It is instructive to briefly summarize here the key features of the main approximation theorem. First, the theorem establishes a non-asymptotic bound between  $Z_n$  and its Gaussian analogue  $\tilde{Z}_n$ . The theorem requires each  $\mathcal{F}_n$  to be pre-Gaussian (i.e., assuming the existence of a version of  $B_n$  that is a tight Gaussian random variable in  $\ell^\infty(\mathcal{F}_n)$ ; see below for the notation), but allows for the case where the “complexity” of  $\mathcal{F}_n$  increases with  $n$ , which places the function classes outside any fixed Donsker class; moreover, neither the process  $\mathbb{G}_n$  nor the supremum statistic  $Z_n$  need to be weakly convergent as  $n \rightarrow \infty$  (even after suitable normalization). Second, the bound in Theorem 2.1 is able to exploit the “local” properties of the class of functions, thereby, when applied to, say, the supremum deviation of kernel type statistics, it leads to tight conditions on the bandwidth for the Gaussian approximation (see the discussion after Theorem 2.1 for details about these features). Note that our bound does not rely on “smoothness” of  $\mathcal{F}_n$  — in contrast, in [52], the bound on the Gaussian approximation for empirical processes depends on the total variation norm of functions. This feature is helpful in deriving good conditions on the number of series functions for the Gaussian approximation of the supremum deviation of projection type statistics treated in Section 3.2 since, for example, the total variation norm is typically large or difficult to control well for such examples. Finally, the theorem only requires finite moments of the envelope function, which should be contrasted with [35, 52, 1, 54] where the classes of functions studied are assumed to be uniformly bounded. Hence the theorem is readily applicable to a wide class of statistical problems to which the previous results are not, at least immediately. We note here that although the bounds we derive are not the sharpest possible in some examples, they are better than previously available bounds in other examples, and are also of interest because of their wide applicability. In fact the results of this paper are already applied in our companion paper [10] and the paper [8] by other authors.

To the best of our knowledge, [48] is the only previous work that considered the problem of directly approximating the distribution of the supremum of the empirical process by that of the corresponding Gaussian process. However, they only cover the case where the class of functions is independent of  $n$  and Donsker as the constant  $C$  in their master Theorem 2 is dependent

on  $\mathcal{F}$  (and how  $C$  depends on  $\mathcal{F}$  is not specified), and their condition (1.4) essentially excludes the case where the “complexity” of  $\mathcal{F}$  grows with  $n$ , which means that their results are not applicable to the statistical problems considered in this paper (see Remark 2.5 or Lemma A.1 ahead). Moreover, their approach is significantly different from ours.

In this paper, we substantially rely on modern empirical process theory. For general references on empirical process theory, we refer to [39, 59, 19, 3]. Section 9.5 of [19] has excellent historical remarks on the Gaussian approximation of empirical processes. For textbook treatments of Yurinskii’s and KMT’s couplings, we refer to [16] and Chapter 10 in [50].

*1.1. Organization.* In Section 2, we present the main approximation theorem (Theorem 2.1). We give a proof of Theorem 2.1 in Section 6. In Section 3, we study applications of Theorem 2.1 to local and series empirical processes arising in nonparametric estimation. Sections 4 and 5 are devoted to developing some technical tools needed to prove Theorem 2.1 and its supporting Lemma 2.2. In Section 4, we prove a new coupling inequality for maxima of sums of random vectors, and in Section 5, we present some inequalities for empirical processes. We put some additional technical proofs, some examples, and additional results in the Appendices. Due to the page limitation, all the Appendices are placed in the Supplemental Material [13].

*1.2. Notation.* Let  $(\Omega, \mathcal{A}, \mathbb{P})$  denote the underlying probability space. We assume that the probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  is rich enough, in the sense that there exists a uniform random variable on  $(0, 1)$  defined on  $(\Omega, \mathcal{A}, \mathbb{P})$  independent of the sample. For a real-valued random variable  $\xi$ , let  $\|\xi\|_q = (\mathbb{E}[|\xi|^q])^{1/q}$ ,  $1 \leq q < \infty$ . For two random variables  $\xi$  and  $\eta$ , we write  $\xi \stackrel{d}{=} \eta$  if they have the same distribution.

For any probability measure  $Q$  on a measurable space  $(S, \mathcal{S})$ , we use the notation  $Qf := \int f dQ$ . Let  $\mathcal{L}^p(Q)$ ,  $p \in [1, \infty]$ , denote the space of all measurable functions  $f : S \rightarrow \mathbb{R}$  such that  $\|f\|_{Q,p} := (Q|f|^p)^{1/p} < \infty$  where  $(Q|f|^p)^{1/p}$  stands for the essential supremum when  $p = \infty$ . We also use the notation  $\|f\|_\infty := \sup_{x \in S} |f(x)|$ . Denote by  $e_Q$  the  $\mathcal{L}^2(Q)$ -semimetric:  $e_Q(f, g) = \|f - g\|_{Q,2}$ ,  $f, g \in \mathcal{L}^2(Q)$ .

For an arbitrary set  $T$ , let  $\ell^\infty(T)$  denote the space of all bounded functions  $T \rightarrow \mathbb{R}$ , equipped with the uniform norm  $\|f\|_T := \sup_{t \in T} |f(t)|$ . We endow  $\ell^\infty(T)$  with the Borel  $\sigma$ -field induced from the norm topology. A random variable in  $\ell^\infty(T)$  refers to a Borel measurable map from  $\Omega$  to  $\ell^\infty(T)$ . For  $\varepsilon > 0$ , an  $\varepsilon$ -net of a semimetric space  $(T, d)$  is a subset  $T_\varepsilon$  of  $T$  such that for every  $t \in T$  there exists a point  $t_\varepsilon \in T_\varepsilon$  with  $d(t, t_\varepsilon) < \varepsilon$ . The  $\varepsilon$ -covering

number  $N(T, d, \varepsilon)$  of  $T$  is the infimum of the cardinality of  $\varepsilon$ -nets of  $T$ , that is,  $N(T, d, \varepsilon) := \inf\{\text{Card}(T_\varepsilon) : T_\varepsilon \text{ is an } \varepsilon\text{-net of } T\}$  (formally define  $N(T, d, 0) := \lim_{\varepsilon \downarrow 0} N(T, d, \varepsilon)$ , where the right limit, possibly being infinite, exists as the map  $\varepsilon \mapsto N(T, d, \varepsilon)$  is non-increasing). For a subset  $A$  of a semimetric space  $(T, d)$ , let  $A^\delta$  denote the  $\delta$ -enlargement of  $A$ , that is,  $A^\delta = \{x \in T : d(x, A) \leq \delta\}$  where  $d(x, A) = \inf_{y \in A} d(x, y)$ .

The standard Euclidean norm is denoted by  $|\cdot|$ . The transpose of a vector  $x$  is denoted by  $x^T$ . We write  $a \lesssim b$  if there exists a universal constant  $C > 0$  such that  $a \leq Cb$ . Unless otherwise stated,  $c, C > 0$  denote universal constants of which the values may change from place to place. For  $a, b \in \mathbb{R}$ , we use the notation  $a \vee b = \max\{a, b\}$  and  $a_+ = a \vee 0$ .

Finally, for a sequence  $\{z_i\}_{i=1}^n$ , we write  $\mathbb{E}_n[z_i] = n^{-1} \sum_{i=1}^n z_i$ , that is,  $\mathbb{E}_n$  abbreviates the symbol  $n^{-1} \sum_{i=1}^n$ . For example,  $\mathbb{E}_n[f(X_i)] = n^{-1} \sum_{i=1}^n f(X_i)$ .

**2. Abstract approximation theorem.** Let  $X_1, \dots, X_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$  with common distribution  $P$ . In all what follows, we assume  $n \geq 3$ . Let  $\mathcal{F}$  be a class of measurable functions  $S \rightarrow \mathbb{R}$ . Here we assume that the class  $\mathcal{F}$  is  $P$ -centered, that is,  $Pf = 0$ ,  $\forall f \in \mathcal{F}$ . This does not lose generality since otherwise we may replace  $\mathcal{F}$  by  $\{f - Pf : f \in \mathcal{F}\}$ . Denote by  $F$  a measurable envelope of  $\mathcal{F}$ , that is,  $F$  is a non-negative measurable function  $S \rightarrow \mathbb{R}$  such that  $F(x) \geq \sup_{f \in \mathcal{F}} |f(x)|$ ,  $\forall x \in S$ .

In this section the sample size  $n$  is fixed, and hence the possible dependence of  $\mathcal{F}$  and  $F$  (and other quantities) on  $n$  is dropped.

We make the following assumptions.

- (A1) The class  $\mathcal{F}$  is *pointwise measurable*, that is, it contains a countable subset  $\mathcal{G}$  such that for every  $f \in \mathcal{F}$  there exists a sequence  $g_m \in \mathcal{G}$  with  $g_m(x) \rightarrow f(x)$  for every  $x \in S$ .
- (A2) For some  $q \geq 2$ ,  $F \in \mathcal{L}^q(P)$ .
- (A3) The class  $\mathcal{F}$  is  $P$ -pre-Gaussian, that is, there exists a tight Gaussian random variable  $G_P$  in  $\ell^\infty(\mathcal{F})$  with mean zero and covariance function

$$\mathbb{E}[G_P(f)G_P(g)] = P(fg) = \mathbb{E}[f(X_1)g(X_1)], \quad \forall f, g \in \mathcal{F}.$$

Assumption (A1) is made to avoid measurability complications. See Section 2.3.1 of [59] for further discussion. This assumption ensures that, for example,  $\sup_{f \in \mathcal{F}} \mathbb{G}_n f = \sup_{f \in \mathcal{G}} \mathbb{G}_n f$ , and hence the former supremum is a measurable map from  $\Omega$  to  $\mathbb{R}$ . Note that by Example 1.5.10 in [59], assumption (A3) implies that  $\mathcal{F}$  is totally bounded for  $e_P$ , and  $G_P$  has sample paths almost surely uniformly  $e_P$ -continuous.

To state the main result, we prepare some notation. For  $\varepsilon > 0$ , define  $\mathcal{F}_\varepsilon = \{f - g : f, g \in \mathcal{F}, e_P(f, g) < \varepsilon \|F\|_{P,2}\}$ . Note that by Theorem 3.1.1 in [19], under assumption (A3), one can extend  $G_P$  to the linear hull of  $\mathcal{F}$  in such a way that  $G_P$  has linear sample paths (recall that the linear hull of  $\mathcal{F}$  is defined as the collection of functions of the form  $\sum_{j=1}^m \alpha_j f_j$  where  $\alpha_j \in \mathbb{R}, f_j \in \mathcal{F}, j = 1, \dots, m$ ). With this in mind, let

$$(3) \quad \phi_n(\varepsilon) = \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}] \vee \mathbb{E}[\|G_P\|_{\mathcal{F}_\varepsilon}].$$

For the notational convenience, let us write

$$(4) \quad H_n(\varepsilon) = \log(N(\mathcal{F}, e_P, \varepsilon \|F\|_{P,2}) \vee n).$$

Note that since  $\mathcal{F}$  is totally bounded for  $e_P$  (because of assumption (A3)),  $H_n(\varepsilon)$  is finite for every  $0 < \varepsilon \leq 1$ . Moreover, write  $M = \max_{1 \leq i \leq n} F(X_i)$  and  $\mathcal{F} \cdot \mathcal{F} = \{fg : f \in \mathcal{F}, g \in \mathcal{F}\}$ . The following is the main theorem of this paper. The proof of the theorem will be given in Section 6.

**THEOREM 2.1 (Gaussian approximation to suprema of empirical processes).** *Suppose that assumptions (A1), (A2) with  $q \geq 3$ , and (A3) are satisfied. Let  $Z = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$ . Let  $\kappa > 0$  be any positive constant such that  $\kappa^3 \geq \mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}]$ . Then for every  $\varepsilon \in (0, 1]$  and  $\gamma \in (0, 1)$ , there exists a random variable  $\tilde{Z} \stackrel{d}{=} \sup_{f \in \mathcal{F}} G_P f$  such that*

$$\mathbb{P}\left\{|Z - \tilde{Z}| > K(q)\Delta_n(\varepsilon, \gamma)\right\} \leq \gamma\{1 + \delta_n(\varepsilon, \gamma)\} + \frac{C \log n}{n},$$

where  $K(q) > 0$  is a constant that depends only on  $q$ , and

$$\begin{aligned} \Delta_n(\varepsilon, \gamma) &:= \phi_n(\varepsilon) + \gamma^{-1/q} \varepsilon \|F\|_{P,2} + n^{-1/2} \gamma^{-1/q} \|M\|_q + n^{-1/2} \gamma^{-2/q} \|M\|_2 \\ &\quad + n^{-1/4} \gamma^{-1/2} (\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F} \cdot \mathcal{F}}])^{1/2} H_n^{1/2}(\varepsilon) + n^{-1/6} \gamma^{-1/3} \kappa H_n^{2/3}(\varepsilon). \\ \delta_n(\varepsilon, \gamma) &:= \frac{1}{4} P\{(F/\kappa)^3 \mathbf{1}(F/\kappa > c\gamma^{-1/3} n^{1/3} H_n(\varepsilon)^{-1/3})\}. \end{aligned}$$

At this point, Theorem 2.1 might seem abstract but in fact it has wide applicability. We provide a general discussion of key features of the theorem in Remark 2.3 below after we present bounds on the main terms in the theorem. See also Corollary 2.2 where we apply Theorem 2.1 to VC type classes where many simplifications of the abstract result are possible.

Recall that we have extended  $G_P$  to the linear hull of  $\mathcal{F}$  in such a way that  $G_P$  has linear sample paths. Hence

$$\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F} \cup (-\mathcal{F})} \mathbb{G}_n f, \quad \|G_P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F} \cup (-\mathcal{F})} G_P f,$$

where  $-\mathcal{F} := \{-f : f \in \mathcal{F}\}$ , from which one can readily deduce the following corollary. Henceforth we only deal with  $\sup_{f \in \mathcal{F}} \mathbb{G}_n f$ .

**COROLLARY 2.1.** *The conclusion of Theorem 2.1 continues to hold with  $Z$  replaced by  $Z = \|\mathbb{G}_n\|_{\mathcal{F}}$ ,  $\tilde{Z}$  replaced by  $\tilde{Z} \stackrel{d}{=} \|\mathbb{G}_P\|_{\mathcal{F}}$ , and with different constants  $K(q)$  and  $C$  where  $K(q)$  depends only on  $q$  and  $C$  is universal.*

Theorem 2.1 is useful only if there are suitable bounds on the following triple of terms, appearing in its statement:

$$(5) \quad \phi_n(\varepsilon), \quad \mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}] \quad \text{and} \quad \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}\cdot\mathcal{F}}].$$

To bound these terms, the entropy method or the more general generic chaining method [58] are useful. We will derive bounds on these terms using the entropy method since typically it leads to readily computable bounds. However, we leave the option of bounding the terms in (5) by other means, e.g., the generic chaining method (in some applications the latter is known to give sharper bounds than the entropy approach).

Consider, as in [59, p.239], the (uniform) entropy integral

$$J(\delta) = J(\delta, \mathcal{F}, F) = \int_0^\delta \sup_Q \sqrt{1 + \log N(\mathcal{F}, e_Q, \varepsilon \|F\|_{Q,2})} d\varepsilon,$$

where the supremum is taken over all finitely discrete probability measures on  $(S, \mathcal{S})$ ; see [59], Sections 2.6 and 2.10.3, and [19], Chapter 4, for examples where the uniform entropy integral can be suitably bounded. We assume the integral is finite:

$$(A4) \quad J(1, \mathcal{F}, F) < \infty.$$

**REMARK 2.1.** In applications  $\mathcal{F}$  and  $F$  (and even  $S$ ) may change with  $n$ , that is,  $\mathcal{F} = \mathcal{F}_n$  and  $F = F_n$ . In that case, assumption (A4) is interpreted as  $J(1, \mathcal{F}_n, F_n) < \infty$  for each  $n$ , but it does allow for the case where  $J(1, \mathcal{F}_n, F_n) \rightarrow \infty$  as  $n \rightarrow \infty$ . ■

We first note the following (standard) fact.

**LEMMA 2.1.** *Assumptions (A2) and (A4) imply assumption (A3).*

For the sake of completeness, we verify this lemma in the Supplemental Material [13]. The following lemma provides bounds on the quantities in (5). Its proof is given in the Supplemental Material [13].

LEMMA 2.2 (Entropy-based bounds on the triple (5)). *Suppose that assumptions (A1), (A2) and (A4) are satisfied. Then for  $\varepsilon \in (0, 1]$ ,*

$$\phi_n(\varepsilon) \lesssim J(\varepsilon)\|F\|_{P,2} + n^{-1/2}\varepsilon^{-2}J^2(\varepsilon)\|M\|_2.$$

Moreover, suppose that assumption (A2) is satisfied with  $q \geq 4$ , and for  $k = 3, 4$ , let  $\delta_k \in (0, 1]$  be any positive constant such that  $\delta_k \geq \sup_{f \in \mathcal{F}} \|f\|_{P,k} / \|F\|_{P,k}$ . Then

$$\begin{aligned} & \mathbb{E}[\|\mathbb{E}_n[|f(X_i)|^3]\|_{\mathcal{F}}] - \sup_{f \in \mathcal{F}} P|f|^3 \\ & \lesssim n^{-1/2}\|M\|_3^{3/2} \left[ J(\delta_3^{3/2}, \mathcal{F}, F)\|F\|_{P,3}^{3/2} + \frac{\|M\|_3^{3/2}J^2(\delta_3^{3/2}, \mathcal{F}, F)}{\sqrt{n}\delta_3^3} \right], \\ & \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F},\mathcal{F}}] \lesssim J(\delta_4^2, \mathcal{F}, F)\|F\|_{P,4}^2 + \frac{\|M\|_4^2J^2(\delta_4^2, \mathcal{F}, F)}{\sqrt{n}\delta_4^4}. \end{aligned}$$

REMARK 2.2 (On the usefulness of the above bounds). The bounds above are designed to handle cases when the suprema of weak moments,  $P|f|^3$  and  $Pf^4$ , are much smaller than the moments of the envelope function, which is the case for all the examples studied in Section 3 where all the proofs for the results in that section follow from application of Corollary 2.2 below, which is a direct consequence of Theorem 2.1 and Lemma 2.2. ■

REMARK 2.3 (**Key features of Theorem 2.1**). Before going to the applications, we discuss the key features of Theorem 2.1. First, Theorem 2.1 does not require uniform boundedness of  $\mathcal{F}$ , and requires only finite moments of the envelope function. This should be contrasted with the fact that many papers working on the Gaussian approximation of empirical processes in the sup-norm, such as [35, 52, 1, 54], required that classes of functions are uniformly bounded. There are, however, many statistical applications where uniform boundedness of the class of functions is too restrictive, and the generality of Theorem 2.1 in this direction will turn out to be useful — a typical example of such an application is the problem of performing inference on a nonparametric regression function with unbounded noise using kernel and series estimation methods. One drawback is that  $\gamma$ , which in applications we take as  $\gamma = \gamma_n \rightarrow 0$ , is typically at most  $O(n^{-1/2})$ , and hence Theorem 2.1 generally gives only “in probability bounds” rather than “almost sure bounds” (though in some cases, it is possible to derive “almost sure bounds” from this theorem; see, in particular, Appendix C of the Supplemental Material). The second feature of Theorem 2.1 is that it is able to exploit the “local” properties of the class of functions  $\mathcal{F}$ . By Lemma 2.2,

typically, we may take  $\kappa^3 \approx \sup_{f \in \mathcal{F}} P|f|^3$  and  $\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}, \mathcal{F}}] \approx \sup_{f \in \mathcal{F}} \sqrt{P f^4}$  (up to logarithmic in  $n$  factors). In some applications, for example, nonparametric kernel and series estimations considered in the next section, the class  $\mathcal{F} = \mathcal{F}_n$  changes with  $n$  and  $\sup_{f \in \mathcal{F}_n} \|f\|_{P,k} / \|F_n\|_{P,k}$  with  $k = 3, 4$  decrease to 0 where  $F_n$  is an envelope function of  $\mathcal{F}_n$ . The bound in Theorem 2.1 (with help of Lemma 2.2) effectively exploits this information and leads to tight conditions on, say, the bandwidth and the number of series functions for the Gaussian approximation; roughly the theorem gives bounds on the approximation error of the form  $(nh_n^d)^{-1/6}$  for kernel estimation and  $(K_n/n)^{-1/6}$  for series estimation (up to logarithmic in  $n$  factors), where  $h_n \rightarrow 0$  is the bandwidth and  $K_n \rightarrow \infty$  is the number of series functions. This feature will be clear from the proofs for the applications in the following section. ■

REMARK 2.4 (An application to VC type classes). Although applications of the general results in this section are not restricted to VC type classes, combination of Theorem 2.1 and Lemma 2.2 will lead to a simple bound for these classes. Recall the definition of VC type classes:

DEFINITION 2.1 (VC type class). Let  $\mathcal{F}$  be a class of measurable functions on a measurable space  $(S, \mathcal{S})$ , to which a measurable envelope  $F$  is attached. We say that  $\mathcal{F}$  is *VC type* with envelope  $F$  if there are constants  $A, v > 0$  such that  $\sup_Q N(\mathcal{F}, e_Q, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^v$  for all  $0 < \varepsilon \leq 1$ , where the supremum is taken over all finitely discrete probability measures on  $(S, \mathcal{S})$ .

Note that the definition of VC type classes allows for unbounded envelopes  $F$ . The VC type class is a wider concept than VC *subgraph* class ([59], Chapter 2.6). The VC type property is “stable” under summation, product, or more generally Lipschitz-type transformations, making it much easier to check whether a function class is VC type; see Lemma A.6 in the Supplemental Material [13].

We have the following corollary of Theorem 2.1, whose proof is given in the Supplemental Material [13].

COROLLARY 2.2 (**Gaussian approximation to suprema of empirical processes indexed by VC type classes**). *Suppose that assumption (A1) is satisfied. In addition, suppose that the class  $\mathcal{F}$  is VC type with an envelope  $F$  and constants  $A \geq e$  and  $v \geq 1$ . Suppose also that for some  $b \geq \sigma > 0$ , and  $q \in [4, \infty]$ , we have  $\sup_{f \in \mathcal{F}} P|f|^k \leq \sigma^2 b^{k-2}$  for  $k = 3, 4$  and  $\|F\|_{P,q} \leq b$ . Let  $Z = \sup_{f \in \mathcal{F}} \mathbb{G}_n f$ . Then for every  $\gamma \in (0, 1)$ , there exist constants  $c, C > 0$*

that depend only on  $q$ , and a random variable  $\tilde{Z} \stackrel{d}{=} \sup_{f \in \mathcal{F}} G_P f$  such that

$$\mathbb{P} \left\{ |Z - \tilde{Z}| > \frac{bK_n}{\gamma^{1/2}n^{1/2-1/q}} + \frac{(b\sigma)^{1/2}K_n^{3/4}}{\gamma^{1/2}n^{1/4}} + \frac{(b\sigma^2K_n^2)^{1/3}}{\gamma^{1/3}n^{1/6}} \right\} \leq C \left( \gamma + \frac{\log n}{n} \right),$$

where  $K_n = cv(\log n \vee \log(Ab/\sigma))$  (“ $1/q$ ” is interpreted as “0” when  $q = \infty$ ).

■

REMARK 2.5 (Gaussian approximation in the Kolmogorov distance). Theorem 2.1 combined with Lemma 2.2 can be used to show that the result (1) holds for some sequence of constants  $r_n \rightarrow 0$  (subject to some conditions; possible rates of  $r_n$  are problem-specific). In statistical applications, however, one is typically interested in the result of the form (here we follow the notation used in Section 1)

$$(6) \quad \sup_{t \in \mathbb{R}} |\mathbb{P}(Z_n \leq t) - \mathbb{P}(\tilde{Z}_n \leq t)| = o(1), \quad n \rightarrow \infty.$$

That is, the approximation of the distribution of  $Z_n$  by that of  $\tilde{Z}_n$  in the Kolmogorov distance is required. To derive (6) from (1), we invoke the following lemma.

LEMMA 2.3 (Gaussian approximation in Kolmogorov distance: non-asymptotic result). *Consider the setting described in the beginning of this section. Suppose that assumptions (A1)-(A3) are satisfied, and that there exist constants  $\underline{\sigma}, \bar{\sigma} > 0$  such that  $\underline{\sigma}^2 \leq Pf^2 \leq \bar{\sigma}^2$  for all  $f \in \mathcal{F}$ . Moreover, suppose that there exist constants  $r_1, r_2 > 0$  and a random variable  $\tilde{Z} \stackrel{d}{=} \sup_{f \in \mathcal{F}} G_P f$  such that  $\mathbb{P}\{|Z - \tilde{Z}| > r_1\} \leq r_2$ . Then*

$$\sup_{t \in \mathbb{R}} |\mathbb{P}(Z \leq t) - \mathbb{P}(\tilde{Z} \leq t)| \leq C_\sigma r_1 \left\{ \mathbb{E}[\tilde{Z}] + \sqrt{1 \vee \log(\bar{\sigma}/r_1)} \right\} + r_2,$$

where  $C_\sigma$  is a constant depending only on  $\underline{\sigma}$  and  $\bar{\sigma}$ .

It is now not difficult to give conditions to deduce (6) from (1). Formally, we state the following lemma.

LEMMA 2.4 (Gaussian approximation in Kolmogorov distance: asymptotic result). *Suppose that there exists a sequence of ( $P$ -centered) classes  $\mathcal{F}_n$  of*

measurable functions  $S \rightarrow \mathbb{R}$  satisfying assumptions (A1)-(A3) with  $\mathcal{F} = \mathcal{F}_n$  for each  $n$ , and that there exist constants  $\underline{\sigma}, \bar{\sigma} > 0$  (independent of  $n$ ) such that  $\underline{\sigma}^2 \leq Pf^2 \leq \bar{\sigma}^2$  for all  $f \in \mathcal{F}_n$ . Let  $Z_n = \sup_{f \in \mathcal{F}_n} \mathbb{G}_n f$ , and denote by  $B_n$  a tight Gaussian random variable in  $\ell^\infty(\mathcal{F}_n)$  with mean zero and covariance function  $\mathbb{E}[B_n(f)B_n(g)] = P(fg)$  for all  $f, g \in \mathcal{F}_n$ . Moreover, suppose that there exist a sequence of random variables  $\tilde{Z}_n \stackrel{d}{=} \sup_{f \in \mathcal{F}_n} B_n f$  and a sequence of constants  $r_n \rightarrow 0$  such that  $|Z_n - \tilde{Z}_n| = O_{\mathbb{P}}(r_n)$  and  $r_n \mathbb{E}[\tilde{Z}_n] = o(1)$  as  $n \rightarrow \infty$ . Then as  $n \rightarrow \infty$ ,  $\sup_{t \in \mathbb{R}} |\mathbb{P}(Z_n \leq t) - \mathbb{P}(\tilde{Z}_n \leq t)| = o(1)$ .

Note here that we allow the case where  $\mathbb{E}[\tilde{Z}_n] \rightarrow \infty$ . In the examples handled in the following section, typically, we have  $\mathbb{E}[\tilde{Z}_n] = O(\sqrt{\log n})$ . We note that the companion work [12] provides multiplier bootstrap methods for uniformly consistent estimation of the map  $t \mapsto \mathbb{P}(\tilde{Z}_n \leq t)$  in applications (see also Remark 3.3 below). ■

**3. Applications.** This section studies applications of Theorem 2.1 and its supporting Lemma 2.2 (via Corollary 2.2) to local and series empirical processes arising in nonparametric estimation via kernel and series methods. In both examples, the classes of functions change with the sample size  $n$  and the corresponding processes  $\mathbb{G}_n$  do not have tight limits. Hence regularity conditions for the Gaussian approximation for the suprema will be of interest. All the proofs in this section, and motivating examples for series empirical processes treated in Section 3.2, are gathered in the Supplemental Material [13].

*3.1. Local empirical processes.* This section applies Theorem 2.1 to the supremum deviation of kernel type statistics. Let  $(Y_1, X_1), \dots, (Y_n, X_n)$  be i.i.d. random variables taking values in the product space  $\mathcal{Y} \times \mathbb{R}^d$ , where  $(\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$  is an arbitrary measurable space. Suppose that there is a class  $\mathcal{G}$  of measurable functions  $\mathcal{Y} \rightarrow \mathbb{R}$ . Let  $k(\cdot)$  be a kernel function on  $\mathbb{R}^d$ . By “kernel function”, we simply mean that  $k(\cdot)$  is integrable with respect to the Lebesgue measure on  $\mathbb{R}^d$  and its integral on  $\mathbb{R}^d$  is normalized to be 1, but we do not assume  $k(\cdot)$  to be non-negative, that is, higher order kernels are allowed. Let  $h_n$  be a sequence of positive constants such that  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ , and let  $\mathcal{I}$  be an arbitrary Borel subset of  $\mathbb{R}^d$ . Consider the kernel-type statistics

$$(7) \quad S_n(x, g) = \frac{1}{nh_n^d} \sum_{i=1}^n g(Y_i) k(h_n^{-1}(X_i - x)), \quad (x, g) \in \mathcal{I} \times \mathcal{G}.$$

Typically, under suitable regularity conditions,  $S_n(x, g)$  will be a consistent estimator of  $\mathbb{E}[g(Y_1) | X_1 = x]p(x)$ , where  $p(\cdot)$  denotes a Lebesgue density of the distribution of  $X_1$  (assuming its existence). For example, when  $g \equiv 1$ ,  $S_n(x, g)$  will be a consistent estimator of  $p(x)$ ; when  $\mathcal{Y} = \mathbb{R}$  and  $g(y) = y$ ,  $S_n(x, g)$  will be a consistent estimator of  $\mathbb{E}[Y_1 | X_1 = x]p(x)$ ; and when  $\mathcal{Y} = \mathbb{R}$  and  $g(\cdot) = 1(\cdot \leq y)$ ,  $y \in \mathbb{R}$ ,  $S_n(x, g)$  will be a consistent estimator of  $\mathbb{P}(Y_1 \leq y | X_1 = x)p(x)$ . In statistical applications, it is often of interest to approximate the distribution of the following quantity:

$$(8) \quad W_n = \sup_{(x, g) \in \mathcal{I} \times \mathcal{G}} c_n(x, g) \sqrt{nh_n^d} (S_n(x, g) - \mathbb{E}[S_n(x, g)]),$$

where  $c_n(x, g)$  is a suitable normalizing constant. A typical choice of  $c_n(x, g)$  would be such that  $\text{Var}(\sqrt{nh_n^d} S_n(x, g)) = c_n(x, g)^{-2} + o(1)$ . Limit theorems for  $W_n$  are developed in [2, 37, 17, 52, 21, 41], among others.

[21] called the process  $g \mapsto \sqrt{nh_n^d} (S_n(x, g) - \mathbb{E}[S_n(x, g)])$  a ‘‘local’’ empirical process at  $x$  (the original definition of the local empirical process in [21] is slightly more general in that  $h_n$  is replaced by a sequence of bi-measurable functions). With a slight abuse of terminology, we also call the process  $(x, g) \mapsto \sqrt{nh_n^d} (S_n(x, g) - \mathbb{E}[S_n(x, g)])$  a local empirical process.

We consider the problem of approximating  $W_n$  by a sequence of suprema of Gaussian processes. For each  $n \geq 1$ , let  $B_n$  be a centered Gaussian process indexed by  $\mathcal{I} \times \mathcal{G}$  with covariance function

$$(9) \quad \begin{aligned} & \mathbb{E}[B_n(x, g)B_n(\check{x}, \check{g})] \\ &= h_n^{-d} c_n(x, g) c_n(\check{x}, \check{g}) \text{Cov}[g(Y_1)k(h_n^{-1}(X_1 - x)), \check{g}(Y_1)k(h_n^{-1}(X_1 - \check{x}))]. \end{aligned}$$

It is expected that under suitable regularity conditions, there is a sequence  $\widetilde{W}_n$  of random variables such that  $\widetilde{W}_n \stackrel{d}{=} \sup_{(x, g) \in \mathcal{I} \times \mathcal{G}} B_n(x, g)$  and as  $n \rightarrow \infty$ ,  $|W_n - \widetilde{W}_n| \xrightarrow{\mathbb{P}} 0$ . We shall argue the validity of this approximation with explicit rates.

We make the following assumptions.

- (B1)  $\mathcal{G}$  is a pointwise measurable class of functions  $\mathcal{Y} \rightarrow \mathbb{R}$  uniformly bounded by a constant  $b > 0$ , and is VC type with envelope  $\equiv b$ .
- (B2)  $k(\cdot)$  is a bounded and continuous kernel function on  $\mathbb{R}^d$ , and such that the class of functions  $\mathcal{K} = \{t \mapsto k(ht + x) : h > 0, x \in \mathbb{R}^d\}$  is VC type with envelope  $\equiv \|k\|_\infty$ .
- (B3) The distribution of  $X_1$  has a bounded Lebesgue density  $p(\cdot)$  on  $\mathbb{R}^d$ .
- (B4)  $h_n \rightarrow 0$  and  $\log(1/h_n) = O(\log n)$  as  $n \rightarrow \infty$ .

(B5)  $C_{\mathcal{I} \times \mathcal{G}} := \sup_{n \geq 1} \sup_{(x,g) \in \mathcal{I} \times \mathcal{G}} |c_n(x,g)| < \infty$ . Moreover, for every fixed  $n \geq 1$  and for every  $(x_m, g_m) \in \mathcal{I} \times \mathcal{G}$  with  $x_m \rightarrow x \in \mathcal{I}$  and  $g_m \rightarrow g \in \mathcal{G}$  pointwise,  $c_n(x_m, g_m) \rightarrow c_n(x, g)$ .

We note that [47] and especially [26, 27] give general sufficient conditions under which  $\mathcal{K}$  is VC type.

We first assume that  $\mathcal{G}$  is uniformly bounded, which will be relaxed later.

**PROPOSITION 3.1** (Gaussian approximation to suprema of local empirical processes: bounded case). *Suppose that assumptions (B1)-(B5) are satisfied. Then for every  $n \geq 1$ , there is a tight Gaussian random variable  $B_n$  in  $\ell^\infty(\mathcal{I} \times \mathcal{G})$  with mean zero and covariance function (9), and there is a sequence  $\widetilde{W}_n$  of random variables such that  $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g) \in \mathcal{I} \times \mathcal{G}} B_n(x,g)$  and as  $n \rightarrow \infty$ ,*

$$|W_n - \widetilde{W}_n| = O_{\mathbb{P}}\{(nh_n^d)^{-1/6} \log n + (nh_n^d)^{-1/4} \log^{5/4} n + (nh_n^d)^{-1/2} \log^{3/2} n\}.$$

Even when  $\mathcal{G}$  is not uniformly bounded, a version of Proposition 3.1 continues to hold provided that suitable restrictions on the moments of the envelope of  $\mathcal{G}$  are assumed. Instead of assumption (B1), we make the following assumption.

(B1)'  $\mathcal{G}$  is a pointwise measurable class of functions  $\mathcal{Y} \rightarrow \mathbb{R}$  with measurable envelope  $G$  such that  $\mathbb{E}[G^q(Y_1)] < \infty$  for some  $q \geq 4$  and  $\sup_{x \in \mathbb{R}^d} \mathbb{E}[G^4(Y_1) | X_1 = x] < \infty$ . Moreover,  $\mathcal{G}$  is VC type with envelope  $G$ .

Then we have the following proposition.

**PROPOSITION 3.2** (Gaussian approximation to suprema of local empirical processes: unbounded case). *Suppose that assumptions (B1)' and (B2)-(B5) are satisfied. Then the conclusion of Proposition 3.1 continues to hold, except for that the speed of approximation is*

$$O_{\mathbb{P}}\{(nh_n^d)^{-1/6} \log n + (nh_n^d)^{-1/4} \log^{5/4} n + (n^{1-2/q} h_n^d)^{-1/2} \log^{3/2} n\}.$$

**REMARK 3.1** (Discussion and comparison to other results). It is instructive to compare Propositions 3.1 and 3.2 with implications of Theorem 1.1 of Rio [52], which is a very sharp result on the Gaussian approximation (in the sup-norm) of general empirical processes indexed by uniformly bounded VC type classes of functions having locally uniformly bounded variation.

1. Rio's [52] Theorem 1.1 is not applicable to the case where the envelope function  $G$  is not bounded. Hence Proposition 3.2 is not covered by [52].

Indeed, we are not aware of any previous result that leads to the conclusion of Proposition 3.2, at least in this generality. For example, [37] considered the Gaussian approximation of  $W_n$  in the case where  $\mathcal{Y} = \mathbb{R}$  and  $g(y) = y$ , but also assumed that the support of  $Y_1$  is bounded. [21] proved in their Theorem 1.1 a weak convergence result for local empirical processes, which, combined with the Skorohod representation and Lemma 4.1 ahead, implies a Gaussian approximation result for  $W_n$  even when  $\mathcal{G}$  is not uniformly bounded (but without explicit rates); however, their Theorem 1.1 (and also Theorem 1.2) is tied with the single value of  $x$ , that is,  $x$  is fixed, since both theorems assume that the “localized” probability measure, localized at a given  $x$ , converges (in a suitable sense) to a fixed probability measure (see assumption (F.ii) in [21]). The same comment applies to [22]. In contrast, our results apply to the case where the supremum is taken over an uncountable set of values of  $x$ , which is relevant to statistical applications such as construction of uniform confidence bands.

2. In the special case of kernel density estimation (i.e.,  $g \equiv 1$ ), Rio’s Theorem 1.1 implies (subject to some regularity conditions) that  $|W_n - \widetilde{W}_n| = O_{a.s.}\{(nh_n^d)^{-1/(2d)}\sqrt{\log n} + (nh_n^d)^{-1/2} \log n\}$  for  $d \geq 2$  (the  $d = 1$  case is formally excluded from [52] but Giné and Nickl showed that the same bound can be obtained for  $d = 1$  case [the proof of Proposition 5 in 28]). Hence Rio-Giné-Nickl’s error rates are better than ours when  $d = 1, 2, 3$ , but ours are better when  $d \geq 4$  (aside from the difference between “in probability” and almost sure bounds). Another approach to couplings of kernel density estimators is proposed in Neumann [45] where the distribution of  $W_n$  is coupled to the distribution of the smoothed bootstrap, which is then coupled to the distribution of the empirical bootstrap. Neumann’s Theorem 3.2 implies that one can construct a sequence  $X_1, \dots, X_n$ , its copy  $\overline{X}_1, \dots, \overline{X}_n$ , and empirical bootstrap sample  $X_1^*, \dots, X_n^*$  from  $\overline{X}_1, \dots, \overline{X}_n$  so that if we define  $W_n^*$  by (7) and (8) with  $X_1, \dots, X_n$  replaced by  $X_1^*, \dots, X_n^*$ , then  $|W_n - W_n^*| = O_{\mathbb{P}}((nh_d)^{-1/(2+d)}(\log n)^{(4+d)/(2(2+d))})$ . Thus Neumann’s error rates of (empirical bootstrap) approximation are better than our error rates of (Gaussian) approximation when  $d \leq 4$  but ours are better when  $d \geq 5$ . Also we note that Neumann’s approach requires similar side conditions as those of Rio’s approach, is tied with kernel density estimation and not as general as ours.

3. Consider, as a second example, kernel regression estimation (that is,  $\mathcal{Y} = \mathbb{R}$  and  $g(y) = y$ ). In order to formally apply Rio’s Theorem 1.1 to this example, we need to assume that, for example,  $(Y_1, X_1)$  is generated in such a way that  $(Y_1, X_1) = (h(U, X_1), X_1)$  where the joint distribution of  $(U, X_1)$  has support  $[0, 1]^{d+1}$  with continuous and positive Lebesgue den-

sity on  $[0, 1]^{d+1}$ , and  $h$  is a function  $[0, 1]^{d+1} \rightarrow \mathbb{R}$  which is bounded and of bounded variation [for example, let  $F_{Y_1|X_1}^{-1}(\cdot | x)$  denote the quantile function of the conditional distribution of  $Y_1$  given  $X_1 = x$  and take  $U$  uniformly distributed on  $(0, 1)$  independent of  $X_1$ ; then  $(Y_1, X_1) \stackrel{d}{=} (F_{Y_1|X_1}^{-1}(U | X_1), X_1)$ , but for the above condition to be met, we need to assume that  $F_{Y_1|X_1}^{-1}(u | x)$  is (bounded and) of bounded variation as a function of  $u$  and  $x$ , which is not a typical assumption in estimation of the conditional mean]. Subject to such side conditions, Rio's Theorem 1.1 leads to the following error rate:  $|W_n - \widetilde{W}_n| = O_{a.s.}\{(n^{d/(d+1)}h_n^d)^{-1/(2d)}\sqrt{\log n} + (nh_n^d)^{-1/2}\log n\}$ . See, for example, [14], Theorem 8. In contrast, Propositions 3.1 and 3.2 do not require such side conditions. Moreover, aside from the difference between “in probability” and almost sure bounds, as long as  $h_n = O(n^{-a})$  for some  $a > 0$ , our error rates are always better when  $d \geq 2$ . When  $d = 1$ , our rate is better as long as  $nh_n^4/\log^c n \rightarrow 0$  (and vice versa) where  $c > 0$  is some constant. ■

REMARK 3.2 (Converting coupling to convergence in Kolmogorov distance). By Remark 2.5, we can convert the results in Propositions 3.1 and 3.2 into convergence of the Kolmogorov distance between the distributions of  $W_n$  and its Gaussian analogue  $\widetilde{W}_n$ . In fact, under either the assumptions of Proposition 3.1 or 3.2, by Dudley's inequality for Gaussian processes [59, Corollary 2.2.8], it is not difficult to deduce that  $\mathbb{E}[\widetilde{W}_n] = O(\sqrt{\log n})$ . Hence if moreover there exists a constant  $\underline{\sigma} > 0$  (independent of  $n$ ) such that  $\text{Var}(c_n(x, g)\sqrt{nh_n^d}S_n(x, g)) \geq \underline{\sigma}^2$  for all  $(x, g) \in \mathcal{I} \times \mathcal{G}$  (giving primitive regularity conditions for this assumption is a standard task; note also that under either the assumptions of Proposition 3.1 or 3.2,  $\text{Var}(c_n(x, g)\sqrt{nh_n^d}S_n(x, g))$  is bounded from above uniformly in  $(x, g) \in \mathcal{I} \times \mathcal{G}$ ), we have

$$|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n) \Rightarrow \sup_{t \in \mathbb{R}} |\mathbb{P}(W_n \leq t) - \mathbb{P}(\widetilde{W}_n \leq t)| = o(1).$$

Note that  $|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n)$  (i) if  $nh_n^d/\log^c n \rightarrow \infty$  under the assumptions of Proposition 3.1, and (ii) if  $n^{(1-2/q)}h_n^d/\log^c n \rightarrow \infty$  under the assumptions of Proposition 3.2, where  $c > 0$  is some constant. These conditions on the bandwidth  $h_n$  are mild, and interestingly they essentially coincide with the conditions on the bandwidth used in establishing exact rates of uniform strong consistency of kernel type estimators in [23, 24]. ■

REMARK 3.3 (Constructing under-smoothed uniform bands). The results in Propositions 3.1 and 3.2 are useful for constructing one- and two-sided uniform confidence bands for various nonparametric functions, such

as density and conditional mean, estimated via kernel methods. For concreteness, consider a kernel density estimator  $\widehat{S}_n(x) = S_n(x, g)$  defined in (7) with  $g \equiv 1$ . Let  $\sigma_n(x) = \sqrt{\text{Var}(\widehat{S}_n(x))}$ , and define  $W_n$  as in (8) with  $c_n(x, g) = 1/(\sigma_n(x)\sqrt{nh_n^d})$ . Also define  $\mathcal{C}_n(x) = [\widehat{S}_n(x) - c(\alpha)\sigma_n(x), \infty)$  where  $c(\alpha)$  is a constant specified later with  $\alpha \in (0, 1)$  a confidence level. Assume that the bandwidth  $h_n$  is chosen in such a way that

$$(10) \quad \sup_{x \in \mathcal{I}} \frac{|\mathbb{E}[\widehat{S}_n(x)] - p(x)|}{\sigma_n(x)} = o(\log^{-1/2} n).$$

Conditions like (10) are typically referred to as under-smoothing [see 28, p.1130 for related discussion]. Then

$$(11) \quad \begin{aligned} \mathbb{P}(p(x) \in \mathcal{C}_n(x), \forall x \in \mathcal{I}) &\leq \mathbb{P}(W_n \leq c(\alpha) + o(\log^{-1/2} n)) \\ &= \mathbb{P}(\widetilde{W}_n \leq c(\alpha) + o(\log^{-1/2} n)) + o(1) = \mathbb{P}(\widetilde{W}_n \leq c(\alpha)) + o(1), \end{aligned}$$

and likewise  $\mathbb{P}(p(x) \in \mathcal{C}_n(x), \forall x \in \mathcal{I}) \geq \mathbb{P}(\widetilde{W}_n \leq c(\alpha)) - o(1)$ , under the conditions specified in Remark 3.2 where  $\widetilde{W}_n$  is defined in Proposition 3.1. Here the last equality in (11) follows from the anti-concentration inequality for Gaussian processes (see Lemma A.1 in the Supplemental Material [13]) together with the fact that  $\mathbb{E}[\widetilde{W}_n] = O(\sqrt{\log n})$ . Hence  $\mathcal{C}_n(\cdot)$  is a one-sided uniform confidence band of level  $\alpha$  if we set  $c(\alpha)$  to be the  $(1 - \alpha)$ -quantile of the distribution of  $\widetilde{W}_n$ , which in turn can be estimated via a bootstrap procedure; see our companion paper [12]. Another way is to use a bound on the  $(1 - \alpha)$ -quantile of  $\widetilde{W}_n$  using sharp deviation inequalities available to Gaussian processes, which leads to analytic construction of confidence bands; see, for example, [14] for this approach. In some applications, the distribution of the approximating Gaussian process is completely known, and in that case the distribution of  $\widetilde{W}_n$  can be simulated via a direct Monte Carlo method; see [53] for such examples. Finally, we mention that there are alternative, yet more conservative, approaches on construction of confidence bands based on non-asymptotic concentration inequalities (and not on Gaussian approximation); see [40] and [33].  $\blacksquare$

**3.2. Series empirical processes.** Here we consider the following problem. Let  $(\eta_1, X_1), \dots, (\eta_n, X_n)$  be i.i.d. random variables taking values in the product space  $\mathcal{E} \times \mathbb{R}^d$ , where  $(\mathcal{E}, \mathcal{A}_{\mathcal{E}})$  is an arbitrary measurable space. Suppose that the support of  $X_1$  is normalized to be  $[0, 1]^d$ , and for each  $K \geq 1$ , there are  $K$  basis functions  $\psi_{K,1}, \dots, \psi_{K,K}$  defined on  $[0, 1]^d$ . Let  $\psi^K(x) = (\psi_{K,1}(x), \dots, \psi_{K,K}(x))^T$ . Examples of such basis functions are

Fourier series, splines, Cohen-Daubechies-Vial (CDV) wavelet bases [15], Hermite polynomials and so on. Let  $K_n$  be a sequence of positive constants such that  $K_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Let  $\mathcal{G}$  be a class of measurable functions  $\mathcal{E} \rightarrow \mathbb{R}$  such that  $\mathbb{E}[g^2(\eta_1)] < \infty$  and  $\mathbb{E}[g(\eta_1) \mid X_1] = 0$  a.s. for all  $g \in \mathcal{G}$ , and let  $\mathcal{I}$  be an arbitrary Borel measurable subset of  $[0, 1]^d$ . Suppose that there are sequences of  $K_n \times K_n$  matrices  $A_{1n}(g)$  and  $A_{2n}(g)$  indexed by  $g \in \mathcal{G}$ . We assume that  $s_{\min}(A_{2n}(g)) > 0$  for all  $g \in \mathcal{G}$ . In what follows, we let  $s_{\min}(A)$  and  $s_{\max}(A)$  denote the minimum and maximum singular values of a matrix  $A$ , respectively. Consider the following empirical process:

$$S_n(x, g) = \frac{\psi^{K_n}(x)^T A_{1n}(g)^T}{|A_{2n}(g)\psi^{K_n}(x)|} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n g(\eta_i) \psi^{K_n}(X_i) \right], \quad x \in \mathcal{I}, g \in \mathcal{G},$$

which we shall call the ‘‘series empirical process’’ (we shall formally follow the convention  $0/0 = 0$ ). The problem here is the Gaussian approximation of the supremum of this series empirical process:

$$W_n := \sup_{(x, g) \in \mathcal{I} \times \mathcal{G}} S_n(x, g).$$

We address this problem in what follows. The study of distributional approximation of this statistic is motivated by inference problems for functions using series (or sieve) estimation. See Examples B.1 and B.2 in the Supplemental Material [13] for concrete examples, coming from nonparametric conditional mean and quantile estimation using the series method. These examples explain and motivate various forms of  $S_n$  arising in mathematical statistics.

Returning to the general setting, let  $B_n$  be a centered Gaussian process indexed by  $\mathcal{I} \times \mathcal{G}$  with covariance function

$$\begin{aligned} & \mathbb{E}[B_n(x, g)B_n(\check{x}, \check{g})] \\ (12) \quad & = \alpha_n(x, g)^T \mathbb{E}[g(\eta_1)\check{g}(\eta_1)\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T] \alpha_n(\check{x}, \check{g}), \end{aligned}$$

where  $\alpha_n(x, g) = A_{1n}(g)\psi^{K_n}(x)/|A_{2n}(g)\psi^{K_n}(x)|$ . It is expected that under suitable regularity conditions, there is a sequence  $\widetilde{W}_n$  of random variables such that  $\widetilde{W}_n \stackrel{d}{=} \sup_{(x, g) \in \mathcal{I} \times \mathcal{G}} B_n(x, g)$  and as  $n \rightarrow \infty$ ,  $|W_n - \widetilde{W}_n| \xrightarrow{\mathbb{P}} 0$ . We shall establish the validity of this approximation with explicit rates.

We make the following assumptions.

- (C1)  $\mathcal{G}$  is a pointwise measurable VC type class of functions  $\mathcal{E} \rightarrow \mathbb{R}$  with measurable envelope  $G$  such that  $\mathbb{E}[g^2(\eta_1)] < \infty$  and  $\mathbb{E}[g(\eta_1) \mid X_1] = 0$  a.s. for all  $g \in \mathcal{G}$ .

- (C2) There exist some constants  $c_1, C_1 > 0$  such that  $s_{\max}(A_{2n}(g)) \leq C_1$  and  $s_{\min}(A_{2n}(g)) \geq c_1$  for all  $g \in \mathcal{G}$  and  $n \geq 1$ .
- (C3)  $\xi_n := \sup_{x \in [0,1]^d} |\psi^{K_n}(x)| \vee 1 < \infty$  and there exists a constant  $C_2 > 0$  such that  $s_{\max}(\mathbb{E}[\psi^{K_n}(X_1)\psi^{K_n}(X_1)^T]) \leq C_2$  for all  $n \geq 1$ . The map  $(x, g) \mapsto A_{1n}(g)\psi^{K_n}(x)/|A_{2n}(g)\psi^{K_n}(x)| =: \alpha_n(x, g)$  is Lipschitz continuous with Lipschitz constant  $\leq L_n(\geq 1)$  in the following sense:

$$(13) \quad \begin{aligned} |\alpha_n(x, g) - \alpha_n(\check{x}, \check{g})| &\leq L_n\{|x - \check{x}| + (\mathbb{E}[(g(\eta_1) - \check{g}(\eta_1))^2])^{1/2}\}, \\ &\forall x, \check{x} \in [0, 1]^d, \forall g, \check{g} \in \mathcal{G}. \end{aligned}$$

Here  $\xi_n$  and  $L_n$  are allowed to diverge as  $n \rightarrow \infty$ .

- (C4)  $\log \xi_n = O(\log n)$  and  $\log L_n = O(\log n)$  as  $n \rightarrow \infty$ .

For many commonly used basis functions such as Fourier series, splines and CDV wavelet bases,  $\xi_n = O(\sqrt{K_n})$  as  $n \rightarrow \infty$ ; see, for example, [31] and [46]. The Lipschitz condition (13) is satisfied if  $\inf_{x \in [0,1]^d} |\psi^{K_n}(x)| \geq c_2 > 0$ ,  $|\psi^{K_n}(x) - \psi^{K_n}(\check{x})| \leq L_{1n}|x - \check{x}|$ , and  $\|A_{1n}(g) - A_{1n}(\check{g})\|_{\text{op}} \vee \|A_{2n}(g) - A_{2n}(\check{g})\|_{\text{op}} \leq L_{2n}(\mathbb{E}[(g(\eta_1) - \check{g}(\eta_1))^2])^{1/2}$ , where  $c_2 > 0$  is a fixed constant and  $L_{1n}, L_{2n}$  are sequences of constants possibly divergent as  $n \rightarrow \infty$  ( $\|A\|_{\text{op}}$  denotes the operator norm of a matrix  $A$ ). Then (13) is satisfied with  $L_n = O(L_{1n} \vee L_{2n})$ . Assumption (C4) states mild growth restrictions on  $K_n$  and  $L_n$ , and is usually satisfied.

**PROPOSITION 3.3** (Gaussian approximation to suprema of series empirical processes). *Suppose that assumptions (C1)-(C4) are satisfied. Moreover, suppose either (i)  $G$  is bounded (i.e.,  $\|G\|_{\infty} < \infty$ ), or (ii)  $\mathbb{E}[G^q(\eta_1)] < \infty$  for some  $q \geq 4$  and  $\sup_{x \in [0,1]^d} \mathbb{E}[G^4(\eta_1) | X_1 = x] < \infty$ . Then for every  $n \geq 1$ , there is a tight Gaussian random variable  $B_n$  in  $\ell^\infty(\mathcal{I} \times \mathcal{G})$  with mean zero and covariance function (12), and there exists a sequence  $\widetilde{W}_n$  of random variables such that  $\widetilde{W}_n \stackrel{d}{=} \sup_{(x,g) \in \mathcal{I} \times \mathcal{G}} B_n(x, g)$  and as  $n \rightarrow \infty$ ,*

$$\begin{aligned} &|W_n - \widetilde{W}_n| \\ &= \begin{cases} O_{\mathbb{P}}\{n^{-1/6}\xi_n^{1/3} \log n + n^{-1/4}\xi_n^{1/2} \log^{5/4} n + n^{-1/2}\xi_n \log^{3/2} n\}, & (i), \\ O_{\mathbb{P}}\{n^{-1/6}\xi_n^{1/3} \log n + n^{-1/4}\xi_n^{1/2} \log^{5/4} n + n^{-1/2+1/q}\xi_n \log^{3/2} n\}, & (ii). \end{cases} \end{aligned}$$

**REMARK 3.4** (Discussion and comparisons with other approximations). Proposition 3.3 is a new result, and its principal attractive feature is the weak requirement on the number of series functions  $K_n$  (recall that, for example, for Fourier series, splines, and CDV wavelet bases, we have  $\xi_n = O(\sqrt{K_n})$ ). Another approach to deduce a result similar to Proposition 3.3

is to apply Yurinskii's coupling (see Theorem 4.2 ahead) to random vectors  $g(\eta_i)\psi^{K_n}(X_i)$ , which, however, requires a rather stringent restriction on  $K_n$ , namely  $K_n^5/n \rightarrow 0$ , for ensuring  $|W_n - \widetilde{W}_n| \xrightarrow{\mathbb{P}} 0$  even in the simplest case where  $\mathcal{E} = \mathbb{R}$  and  $g(\eta) = \eta$ . See, for example, [14], Theorem 7. Moreover, the use of Rio's [52] Theorem 1.1 here is not effective since the total variation bound is large or difficult to control well in this example, which results in restrictive conditions on  $K_n$  (also Rio's [52] Theorem 1.1 does not cover case (ii) where  $G$  may not be bounded). ■

REMARK 3.5 (Converting coupling to convergence in Kolmogorov distance). As before, we can convert the results in Proposition 3.3 into convergence of the Kolmogorov distance between the distributions of  $W_n$  and its Gaussian analogue  $\widetilde{W}_n$ . Suppose that  $\xi_n = O(\sqrt{K_n})$ . By Dudley's inequality for Gaussian processes [59, Corollary 2.2.8], it is not difficult to deduce that  $\mathbb{E}[\widetilde{W}_n] = O(\sqrt{\log n})$  under the assumptions of Proposition 3.3. Hence if moreover there exists a constant  $\underline{\sigma} > 0$  (independent of  $n$ ) such that  $\text{Var}(S_n(x, g)) \geq \underline{\sigma}^2$  for all  $(x, g) \in \mathcal{I} \times \mathcal{G}$ , by Lemma 2.4, we have

$$|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n) \Rightarrow \sup_{t \in \mathbb{R}} |\mathbb{P}(W_n \leq t) - \mathbb{P}(\widetilde{W}_n \leq t)| = o(1).$$

Note that  $|W_n - \widetilde{W}_n| = o_{\mathbb{P}}(\log^{-1/2} n)$  if  $K_n(\log n)^c/n \rightarrow 0$  in case (i) and if  $K_n(\log n)^c/n^{1-2/q} \rightarrow 0$  in case (ii), where  $c > 0$  is some constant. These requirements on  $K_n$  are mild, in view of the fact that at least  $K_n/n \rightarrow 0$  is needed for consistency (in the  $L^2$ -norm) of the series estimator [see 32]. ■

REMARK 3.6 (Constructing under-smoothed uniform confidence bands). Results in Proposition 3.3 can be used for constructing one- and two-sided uniform confidence bands for various nonparametric functions, such as density, conditional mean, and conditional quantile, estimated via series methods following the same arguments as those described in Remark 3.3 above. ■

#### 4. A coupling inequality for maxima of sums of random vectors.

The main ingredient in the proof of Theorem 2.1 is a new coupling inequality for maxima of sums of random vectors, which is stated below.

THEOREM 4.1 (A coupling inequality for maxima of sums of random vectors). *Let  $X_1, \dots, X_n$  be independent random vectors in  $\mathbb{R}^p$  with mean zero and finite absolute third moments, that is,  $\mathbb{E}[X_{ij}] = 0$  and  $\mathbb{E}[|X_{ij}|^3] < \infty$  for all  $1 \leq i \leq n$  and  $1 \leq j \leq p$ . Consider the statistic  $Z = \max_{1 \leq j \leq p} \sum_{i=1}^n X_{ij}$ .*

Let  $Y_1, \dots, Y_n$  be independent random vectors in  $\mathbb{R}^p$  with  $Y_i \sim N(0, \mathbb{E}[X_i X_i^T])$ ,  $1 \leq i \leq n$ . Then for every  $\beta > 0$  and  $\delta > 1/\beta$ , there exists a random variable  $\tilde{Z} \stackrel{d}{=} \max_{1 \leq j \leq p} \sum_{i=1}^n Y_{ij}$  such that

$$\mathbb{P}(|Z - \tilde{Z}| > 2\beta^{-1} \log p + 3\delta) \leq \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon},$$

where  $\varepsilon = \varepsilon_{\beta, \delta}$  is given by

$$(14) \quad \varepsilon = \sqrt{e^{-\alpha}(1 + \alpha)} < 1, \quad \alpha = \beta^2 \delta^2 - 1 > 0,$$

and

$$\begin{aligned} B_1 &= \mathbb{E} \left[ \max_{1 \leq j, k \leq p} \left| \sum_{i=1}^n (X_{ij} X_{ik} - \mathbb{E}[X_{ij} X_{ik}]) \right| \right], \\ B_2 &= \mathbb{E} \left[ \max_{1 \leq j \leq p} \sum_{i=1}^n |X_{ij}|^3 \right], \\ B_3 &= \sum_{i=1}^n \mathbb{E} \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \cdot 1 \left( \max_{1 \leq j \leq p} |X_{ij}| > \beta^{-1}/2 \right) \right]. \end{aligned}$$

A different, though related, Gaussian approximation inequality was obtained in Theorem 2.1 of [11] with different techniques. We have chosen to present a new theorem here because 1) it is based on the Stein's exchangeable pairs technique, which is well understood in the literature, and our theorem might be helpful for deriving further results in the future; 2) applying Theorem 2.1 of [11] here would require solving a complicated optimization problem to find the best bound for the coupling problem; and 3) our new theorem does not require truncating normal random vectors, allowing us to avoid an additional layer of complication in the final application to empirical processes.

The following corollary is useful for many applications. Recall  $n \geq 3$ .

**COROLLARY 4.1** (An applied coupling inequality for maxima of sums of random vectors). *Consider the same setup as in Theorem 4.1. Then for every  $\delta > 0$ , there exists a random variable  $\tilde{Z} \stackrel{d}{=} \max_{1 \leq j \leq p} \sum_{i=1}^n Y_{ij}$  such that*

$$(15) \quad \mathbb{P}(|Z - \tilde{Z}| > 16\delta) \lesssim \delta^{-2} \{B_1 + \delta^{-1}(B_2 + B_4) \log(p \vee n)\} \log(p \vee n) + \frac{\log n}{n},$$

where  $B_1$  and  $B_2$  are as in Theorem 4.1, and

$$B_4 = \sum_{i=1}^n \mathbb{E} \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \cdot 1 \left( \max_{1 \leq j \leq p} |X_{ij}| > \delta / \log(p \vee n) \right) \right].$$

PROOF OF COROLLARY 4.1. In Theorem 4.1, take  $\beta = 2\delta^{-1} \log(p \vee n)$ . Then  $\alpha = \beta^2 \delta^2 - 1 = 4 \log^2(p \vee n) - 1 \geq 2 \log(p \vee n)$  (recall  $n \geq 3 > e$ ), so that  $\varepsilon \leq 2 \log(p \vee n)/(p \vee n) \leq 2n^{-1} \log n$ . This completes the proof. ■

Theorem 4.1 is a coupling inequality similar in nature to Yurinskii's [61] coupling for sums of random vectors (as opposed to the maxima of such vectors as in the current theorem). Before proving Theorem 4.1, let us first recall Yurinskii's coupling inequality.

THEOREM 4.2 (Yurinskii's coupling for sums of random vectors; [61]; see also [38]). *Consider the same setup as in Theorem 4.1. Let  $S_n = \sum_{i=1}^n X_i$ . Then for every  $\delta > 0$ , there exists a random vector  $T_n \stackrel{d}{=} \sum_{i=1}^n Y_i$  such that*

$$\mathbb{P}(|S_n - T_n| > 3\delta) \lesssim B_0 \left( 1 + \frac{|\log(1/B_0)|}{p} \right),$$

where  $B_0 = p\delta^{-3} \sum_{i=1}^n \mathbb{E}[|X_i|^3]$ .

For the proof, see [50], Section 10.4. Because of the general fact that  $\max_{1 \leq j \leq n} |x_j| \leq |x|$  for  $x \in \mathbb{R}^p$ , one has

$$\left| \max_{1 \leq j \leq p} (S_n)_j - \max_{1 \leq j \leq n} (T_n)_j \right| \leq \max_{1 \leq j \leq p} |(S_n - T_n)_j| \leq |S_n - T_n|.$$

Hence if we take  $\tilde{Z} = \max_{1 \leq j \leq p} (T_n)_j$ ,

$$(16) \quad \mathbb{P}(|Z - \tilde{Z}| > 3\delta) \lesssim B_0 \left( 1 + \frac{|\log(1/B_0)|}{p} \right).$$

Unfortunately, when  $p$  is large, the right side needs not be small. This is because  $B_0$  is proportional to  $\sum_{i=1}^n \mathbb{E}[|X_i|^3]$  and this quantity may be larger than what we want.

To better understand the difference between (15) and (16), consider the situation where  $p$  is indexed by  $n$  and  $p = p_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Moreover, consider the simple case where  $X_{ij} = x_{ij}/\sqrt{n}$  and  $|x_{ij}| \leq b$  ( $x_{ij}$  are random;  $b$  is a fixed constant). Then  $B_1 = O(n^{-1/2} \log^{1/2} p_n)$ ,  $B_2 + B_4 = O(n^{-1/2})$ . The former estimate is deduced from the fact that, using the symmetrization and the maximal inequality for Rademacher averages conditional on  $X_1, \dots, X_n$  [use 59, Lemmas 2.2.2 and 2.2.7], one has  $B_1 \lesssim \sqrt{\log(1+p)} \mathbb{E}[\max_{1 \leq j \leq p} (\sum_{i=1}^n X_{ij}^4)^{1/2}]$ . On the other hand,  $p_n \sum_{i=1}^n |X_i|^3 = O(n^{-1/2} p_n^{5/2})$ . Therefore, to make  $|Z - \tilde{Z}| \xrightarrow{\mathbb{P}} 0$ , the former (15) allows  $p_n$  to be of an exponential order ( $p_n$  can be as large as  $\log p_n = o(n^{1/4})$ ); hence,

for example,  $p_n$  can be of order  $e^{n^\alpha}$  for  $0 < \alpha < 1/4$ ), while the latter (16) restricts  $p_n$  to be  $p_n = o(n^{1/5})$ . Note that, under the exponential moment condition, instead of Yurinskii's coupling, we can use Zaitsev's coupling inequality [62, Theorem 1.1] but it still requires  $p_n = o(n^{1/5})$  to deduce that  $|Z - \tilde{Z}| \xrightarrow{\mathbb{P}} 0$  (although by using Zaitsev's coupling, we indeed have an exponential type inequality for  $|Z - \tilde{Z}|$ ).

**REMARK 4.1** (Connection to Theorem 2.1). The importance of Theorem 4.1 in the context of the proof of Theorem 2.1 is described as follows. In the proof of Theorem 2.1, we make a finite approximation of  $\mathcal{F}$  by a minimal  $\varepsilon\|F\|_{P,2}$ -net of  $(\mathcal{F}, e_P)$  and apply Theorem 4.1 to the “discretized” empirical process; hence in this application,  $p = N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$ . The fact that Theorem 4.1 allows for “large”  $p$  means that a “finer” discretization is possible, and as a result, the bound in Theorem 2.1 depends on the covering number  $N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$  only through its logarithm:  $\log N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$ . ■

We will use a version of Strassen's theorem to prove Theorem 4.1. We state it for the reader's convenience. The proof of this result can be found in the Supplemental Material [13].

**LEMMA 4.1** (An implication of Strassen's theorem). *Let  $\mu$  and  $\nu$  be Borel probability measures on  $\mathbb{R}$ , and let  $V$  be a random variable defined on a probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  with distribution  $\mu$ . Suppose that the probability space  $(\Omega, \mathcal{A}, \mathbb{P})$  admits a uniform random variable on  $(0, 1)$  independent of  $V$ . Let  $\varepsilon > 0$  and  $\delta > 0$  be two positive constants. Then there exists a random variable  $W$ , defined on  $(\Omega, \mathcal{A}, \mathbb{P})$ , with distribution  $\nu$  such that  $\mathbb{P}(|V - W| > \delta) \leq \varepsilon$  if and only if  $\mu(A) \leq \nu(A^\delta) + \varepsilon$  for every Borel subset  $A$  of  $\mathbb{R}$ .*

**PROOF OF THEOREM 4.1.** For the notational convenience, write  $e_\beta = \beta^{-1} \log p$ . Construct  $Y_1, \dots, Y_n$  independent of  $X_1, \dots, X_n$ . By Lemma 4.1, the conclusion follows if we can prove that for every Borel subset  $A$  of  $\mathbb{R}$ ,

$$\mathbb{P}(Z \in A) \leq \mathbb{P}(\tilde{Z}^* \in A^{2e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1}\{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon},$$

where  $\tilde{Z}^* := \max_{1 \leq j \leq p} \sum_{i=1}^n Y_{ij}$ . Let  $S_n = \sum_{i=1}^n X_i$  and  $T_n = \sum_{i=1}^n Y_i$ . Fix any Borel subset  $A$  of  $\mathbb{R}$ . We divide the proof into several steps.

**Step 1:** We approximate the non-smooth map  $x \mapsto 1_A(\max_{1 \leq j \leq p} x_j)$  by a smooth function. The first step is to approximate the map  $x \mapsto \max_{1 \leq j \leq p} x_j$  by a smooth function. Consider the function  $F_\beta : \mathbb{R}^p \rightarrow \mathbb{R}$  defined by  $F_\beta(x) = \beta^{-1} \log(\sum_{j=1}^p e^{\beta x_j})$ , which gives a smooth approximation of  $\max_{1 \leq j \leq p} x_j$ ;

this function arises in definition of free energy in spin glasses [49]. Indeed, an elementary calculation gives the following inequality: for every  $x = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ ,

$$(17) \quad \max_{1 \leq j \leq p} x_j \leq F_\beta(x) \leq \max_{1 \leq j \leq p} x_j + \beta^{-1} \log p.$$

See [6]. Hence we have

$$\mathbb{P}(Z \in A) \leq \mathbb{P}(F_\beta(S_n) \in A^{e\beta}) = \mathbb{E}[1_{A^{e\beta}}(F_\beta(S_n))].$$

**Step 2:** The next step is to approximate the indicator function  $t \mapsto 1_A(t)$  by a smooth function. This step is rather standard.

**LEMMA 4.2.** *Let  $\beta > 0$  and  $\delta > 1/\beta$ . For every Borel subset  $A$  of  $\mathbb{R}$ , there exists a smooth function  $g : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\|g'\|_\infty \leq \delta^{-1}$ ,  $\|g''\|_\infty \leq C\beta\delta^{-1}$ ,  $\|g'''\|_\infty \leq C\beta^2\delta^{-1}$ , and*

$$(1 - \varepsilon)1_A(t) \leq g(t) \leq \varepsilon + (1 - \varepsilon)1_{A^{3\delta}}(t), \quad \forall t \in \mathbb{R},$$

where  $\varepsilon = \varepsilon_{\beta, \delta}$  is given by (14).

**PROOF OF LEMMA 4.2.** The proof is due to [50], Lemma 10.18 (p. 248). Let  $\rho(\cdot, \cdot)$  denote the Euclidean distance on  $\mathbb{R}$ . Then consider the function  $h(t) = (1 - \rho(t, A^\delta)/\delta)_+$ . Note that  $h$  is Lipschitz continuous with Lipschitz constant  $\leq \delta^{-1}$ . Construct a smooth approximation of  $h(t)$  by

$$g(t) = \frac{\beta}{\sqrt{2\pi}} \int_{\mathbb{R}} h(s) e^{-\frac{1}{2}\beta^2(s-t)^2} ds = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(t + \beta^{-1}z) e^{-\frac{1}{2}z^2} dz.$$

Then the map  $t \mapsto g(t)$  is infinitely differentiable, and

$$\|g'\|_\infty \leq \delta^{-1}, \quad \|g''\|_\infty \leq C\beta\delta^{-1}, \quad \|g'''\|_\infty \leq C\beta^2\delta^{-1}.$$

The rest of the proof is the same as [50], Lemma 10.18 and omitted.  $\blacksquare$

Apply Lemma 4.2 to  $A = A^{e\beta}$  to construct a suitable function  $g$ . Then

$$\mathbb{E}[1_{A^{e\beta}}(F_\beta(S_n))] \leq (1 - \varepsilon)^{-1} \mathbb{E}[g \circ F_\beta(S_n)].$$

**Step 3:** The next step uses Stein's method to compare  $\mathbb{E}[g \circ F_\beta(S_n)]$  and  $\mathbb{E}[g \circ F_\beta(T_n)]$ . The following argument is inspired by [7], Theorem 7. We first make some complimentary computations. Here for a smooth function  $f : \mathbb{R}^p \rightarrow \mathbb{R}$ , we use the notation  $\partial_j f(x) = \partial f(x)/\partial x_j$ ,  $\partial_j \partial_k f(x) = \partial^2 f(x)/\partial x_j \partial x_k$ , and so on.

LEMMA 4.3. *Let  $\beta > 0$ . For every  $g \in C^3(\mathbb{R})$ ,*

$$(18) \quad \sum_{j,k=1}^p |\partial_j \partial_k (g \circ F_\beta)(x)| \leq \|g''\|_\infty + 2\|g'\|_\infty \beta,$$

$$(19) \quad \sum_{j,k,l=1}^p |\partial_j \partial_k \partial_l (g \circ F_\beta)(x)| \leq \|g'''\|_\infty + 6\|g''\|_\infty \beta + 6\|g'\|_\infty \beta^2.$$

Moreover, let  $U_{jkl}(x) := \sup\{|\partial_j \partial_k \partial_l (g \circ F_\beta)(x+y)| : y \in \mathbb{R}^p, |y_j| \leq \beta^{-1}, 1 \leq \forall j \leq p\}$ . Then

$$(20) \quad \sum_{j,k,l=1}^p U_{jkl}(x) \leq C(\|g'''\|_\infty + \|g''\|_\infty \beta + \|g'\|_\infty \beta^2).$$

PROOF OF LEMMA 4.3. Let  $\delta_{jk} = 1(j = k)$ . A direct calculation gives

$$\partial_j F_\beta(x) = \pi_j(z), \quad \partial_j \partial_k F_\beta(x) = \beta w_{jk}(x), \quad \partial_j \partial_k \partial_l F_\beta(x) = \beta^2 q_{jkl}(x),$$

where

$$\begin{aligned} \pi_j(x) &= e^{\beta x_j} / \sum_{k=1}^p e^{\beta x_k}, \quad w_{jk}(x) = (\pi_j \delta_{jk} - \pi_j \pi_k)(x), \\ q_{jkl}(x) &= (\pi_j \delta_{jl} \delta_{jk} - \pi_j \pi_l \delta_{jk} - \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(x). \end{aligned}$$

By these expressions, we have

$$\pi_j(x) \geq 0, \quad \sum_{j=1}^p \pi_j(x) = 1, \quad \sum_{j,k=1}^p |w_{jk}(x)| \leq 2, \quad \sum_{j,k,l=1}^p |q_{jkl}(x)| \leq 6.$$

Inequalities (18) and (19) follow from these relations and the following computation.

$$\begin{aligned} \partial_j (g \circ F_\beta)(x) &= (g' \circ F_\beta)(x) \pi_j(x), \\ \partial_j \partial_k (g \circ F_\beta)(x) &= (g'' \circ F_\beta)(x) \pi_j(x) \pi_k(x) + (g' \circ F_\beta)(x) \beta w_{jk}(x), \\ \partial_j \partial_k \partial_l (g \circ F_\beta)(x) &= (g''' \circ F_\beta)(x) \pi_j(x) \pi_k(x) \pi_l(x) \\ &\quad + (g'' \circ F_\beta)(x) \beta (w_{jk}(x) \pi_l(x) + w_{jl}(x) \pi_k(x) + w_{kl}(x) \pi_j(x)) \\ &\quad + (g' \circ F_\beta)(x) \beta^2 q_{jkl}(x). \end{aligned}$$

For the last inequality (20), it is standard to see that whenever  $|y_j| \leq \beta^{-1}, 1 \leq \forall j \leq p$ , we have  $\pi_j(x+y) \leq e^2 \pi_j(x)$ , from which the desired inequality follows.  $\blacksquare$

For  $i = 1, \dots, n$ , let  $X'_i$  be an independent copy of  $X_i$ . Let  $I$  be a uniform random variable on  $\{1, \dots, n\}$  independent of all the other variables. Define  $S'_n := S_n - X_I + X'_I$ . For  $\lambda \in \mathbb{R}^p$ ,

$$\begin{aligned} \mathbb{E}[e^{\sqrt{-1}\lambda^T S'_n}] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[e^{\sqrt{-1}\lambda^T (S_n - X_i)}] \mathbb{E}[e^{\sqrt{-1}\lambda^T X'_i}] \\ &= \frac{1}{n} \sum_{i=1}^n \prod_{j \neq i} \mathbb{E}[e^{\sqrt{-1}\lambda^T X_j}] \mathbb{E}[e^{\sqrt{-1}\lambda^T X_i}] = \prod_{i=1}^n \mathbb{E}[e^{\sqrt{-1}\lambda^T X_i}] = \mathbb{E}[e^{\sqrt{-1}\lambda^T S_n}]. \end{aligned}$$

Hence  $S'_n \stackrel{d}{=} S_n$ . Also with  $X_1^n = \{X_1, \dots, X_n\}$ ,

$$(21) \quad \mathbb{E}[S'_n - S_n \mid X_1^n] = \mathbb{E}[X'_I - X_I \mid X_1^n] = -n^{-1} S_n,$$

and

$$\begin{aligned} \mathbb{E}[(S'_n - S_n)(S'_n - S_n)^T \mid X_1^n] &= \mathbb{E}[(X'_I - X_I)(X'_I - X_I)^T \mid X_1^n] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(X'_i - X_i)(X'_i - X_i)^T \mid X_1^n] = \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[X_i X_i^T] + X_i X_i^T) \\ &= \frac{2}{n} \sum_{i=1}^n \mathbb{E}[X_i X_i^T] + \frac{1}{n} \sum_{i=1}^n (X_i X_i^T - \mathbb{E}[X_i X_i^T]) \\ (22) \quad &= \frac{2}{n} \sum_{i=1}^n \mathbb{E}[X_i X_i^T] + n^{-1} V, \end{aligned}$$

where  $V$  is the  $p \times p$  matrix defined by  $V = (V_{jk})_{1 \leq j, k \leq p} = \sum_{i=1}^n (X_i X_i^T - \mathbb{E}[X_i X_i^T])$ .

For the notational convenience, write  $f = g \circ F_\beta$ . Consider

$$h(x) = \int_0^1 \frac{1}{2t} \mathbb{E}[f(\sqrt{t}x + \sqrt{1-t}T_n) - f(T_n)] dt.$$

Then Lemma 1 of [44] implies

$$\sum_{j=1}^p x_j \partial_j h(x) - \sum_{j,k=1}^p \sum_{i=1}^n \mathbb{E}[X_{ij} X_{ik}] \partial_j \partial_k h(x) = f(x) - \mathbb{E}[f(T_n)],$$

and especially

$$(23) \quad \begin{aligned} \mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)] &= \mathbb{E} \left[ \sum_{j=1}^p \sum_{i=1}^n X_{ij} \partial_j h(S_n) \right] \\ &\quad - \mathbb{E} \left[ \sum_{j,k=1}^p \sum_{i=1}^n \mathbb{E}[X_{ij} X_{ik}] \partial_j \partial_k h(S_n) \right]. \end{aligned}$$

Denote by  $\nabla h(x)$  and  $\text{Hess } h(x)$  the gradient vector and the Hessian matrix of  $h(x)$ , respectively. Let

$$R = h(S'_n) - h(S_n) - (S'_n - S_n)^T \nabla h(S_n) - 2^{-1} (S'_n - S_n)^T (\text{Hess } h(S_n)) (S'_n - S_n).$$

Then one has

$$\begin{aligned} 0 &= n\mathbb{E}[h(S'_n) - h(S_n)] \quad (\text{as } S'_n \stackrel{d}{=} S_n) \\ &= n\mathbb{E}[(S'_n - S_n)^T \nabla h(S_n) + 2^{-1} (S'_n - S_n)^T (\text{Hess } h(S_n)) (S'_n - S_n) + R] \\ &= n\mathbb{E}\left[\mathbb{E}[(S'_n - S_n)^T \mid X_1^n] \nabla h(S_n) + 2^{-1} \text{Tr}\left((\text{Hess } h(S_n)) \mathbb{E}[(S'_n - S_n)(S'_n - S_n)^T \mid X_1^n]\right) + R\right] \\ &= \mathbb{E}\left[-\sum_{j=1}^p \sum_{i=1}^n X_{ij} \partial_j h(S_n) + \sum_{j,k=1}^p \sum_{i=1}^n \mathbb{E}[X_{ij} X_{ik}] \partial_j \partial_k h(S_n)\right] \\ &\quad + \mathbb{E}\left[\frac{1}{2} \sum_{j,k=1}^p V_{jk} \partial_j \partial_k h(S_n) + nR\right] \quad (\text{by (21) and (22)}) \\ &= -\mathbb{E}[f(S_n)] + \mathbb{E}[f(T_n)] + \mathbb{E}\left[\frac{1}{2} \sum_{j,k=1}^p V_{jk} \partial_j \partial_k h(S_n) + nR\right], \quad (\text{by (23)}) \end{aligned}$$

that is,

$$\mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)] = \mathbb{E}\left[\frac{1}{2} \sum_{j,k=1}^p V_{jk} \partial_j \partial_k h(S_n) + nR\right].$$

Using Lemma 4.3, one has

$$\left| \sum_{j,k=1}^p V_{jk} \partial_j \partial_k h(S_n) \right| \leq \max_{1 \leq j,k \leq p} |V_{jk}| \sum_{j,k=1}^p |\partial_j \partial_k h(S_n)| \leq C\beta\delta^{-1} \max_{1 \leq j,k \leq p} |V_{jk}|,$$

and with  $\Delta_i := (\Delta_{i1}, \dots, \Delta_{ip})^T := X'_i - X_i$ ,

$$\begin{aligned} |\mathbb{E}[nR]| &= \left| \mathbb{E}\left[\frac{1}{2} \sum_{i=1}^n \sum_{j,k,l=1}^p \Delta_{ij} \Delta_{ik} \Delta_{il} (1-\theta)^2 \partial_j \partial_k \partial_l h(S_n + \theta \Delta_i)\right] \right| \\ &\quad (\theta \sim U(0,1) \text{ independent of all the other variables}) \\ (24) \quad &\leq \frac{1}{2} \mathbb{E}\left[\sum_{i=1}^n \sum_{j,k,l=1}^p |\Delta_{ij} \Delta_{ik} \Delta_{il}| \cdot |\partial_j \partial_k \partial_l h(S_n + \theta \Delta_i)|\right]. \end{aligned}$$

Let  $\chi_i = 1(\max_{1 \leq j \leq p} |\Delta_{ij}| \leq \beta^{-1})$  and  $\chi_i^c := 1 - \chi_i$ . Then

$$(24) = \frac{1}{2} \mathbb{E} \left[ \sum_{i=1}^n \chi_i^* \right] + \frac{1}{2} \mathbb{E} \left[ \sum_{i=1}^n \chi_i^{c*} \right] =: \frac{1}{2} [(A) + (B)].$$

Observe that

$$\begin{aligned} (A) &\leq \mathbb{E} \left[ \sum_{j,k,l=1}^p \max_{1 \leq i \leq n} (\chi_i \cdot |\partial_j \partial_k \partial_l h(S_n + \theta \Delta_i)|) \times \max_{1 \leq j,k,l \leq p} \sum_{i=1}^n |\Delta_{ij} \Delta_{ik} \Delta_{il}| \right] \\ &\leq C \beta^2 \delta^{-1} \mathbb{E} \left[ \max_{1 \leq j,k,l \leq p} \sum_{i=1}^n |\Delta_{ij} \Delta_{ik} \Delta_{il}| \right] \quad (\text{by (20)}) \\ &\leq C \beta^2 \delta^{-1} \mathbb{E} \left[ \max_{1 \leq j \leq p} \sum_{i=1}^n |\Delta_{ij}|^3 \right] \leq C \beta^2 \delta^{-1} \mathbb{E} \left[ \max_{1 \leq j \leq p} \sum_{i=1}^n |X_{ij}|^3 \right] = C \beta^2 \delta^{-1} B_2, \end{aligned}$$

and

$$\begin{aligned} (B) &\leq C \beta^2 \delta^{-1} \sum_{i=1}^n \mathbb{E} \left[ \chi_i^c \max_{1 \leq j \leq p} |\Delta_{ij}|^3 \right] \quad (\text{by (19)}) \\ &\leq C \beta^2 \delta^{-1} \sum_{i=1}^n \mathbb{E} \left[ \chi_i^c \max_{1 \leq j \leq p} |X_{ij}|^3 \right]. \quad (\text{by symmetry}) \end{aligned}$$

As  $\chi_i^c \leq 1(\max_{1 \leq j \leq p} |X_{ij}| > \beta^{-1}/2) + 1(\max_{1 \leq j \leq p} |X'_{ij}| > \beta^{-1}/2)$ , we have

$$\begin{aligned} \mathbb{E} \left[ \chi_i^c \max_{1 \leq j \leq p} |X_{ij}|^3 \right] &\leq \mathbb{E} \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \cdot 1 \left( \max_{1 \leq j \leq p} |X_{ij}| > \beta^{-1}/2 \right) \right] \\ (25) \quad &+ \mathbb{E} \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \right] \cdot \mathbb{P} \left( \max_{1 \leq j \leq p} |X_{ij}| > \beta^{-1}/2 \right). \end{aligned}$$

We here recall Chebyshev's association inequalities stated in the following lemma. For a proof, see, for example, Theorem 2.14 in [3].

**LEMMA 4.4** (Chebyshev's association inequalities). *Let  $\varphi$  and  $\psi$  be functions defined on an interval  $\mathcal{I}$  in  $\mathbb{R}$ , and let  $\xi$  be a random variable such that  $\mathbb{P}(\xi \in \mathcal{I}) = 1$ . Suppose that  $\mathbb{E}[|\varphi(\xi)|] < \infty$ ,  $\mathbb{E}[|\psi(\xi)|] < \infty$  and  $\mathbb{E}[|\varphi(\xi)\psi(\xi)|] < \infty$ . Then  $\text{Cov}(\varphi(\xi), \psi(\xi)) \geq 0$  if  $\varphi$  and  $\psi$  are monotone in the same direction, and  $\text{Cov}(\varphi(\xi), \psi(\xi)) \leq 0$  if  $\varphi$  and  $\psi$  are monotone in the opposite direction.*

Since the maps  $t \mapsto t^3$  and  $t \mapsto 1(t > \beta^{-1}/2)$  are non-decreasing on  $[0, \infty)$ , the second term on the right side of (25) is not larger than the first term. Hence

$$(B) \leq C\beta^2\delta^{-1} \sum_{i=1}^n \mathbb{E} \left[ \max_{1 \leq j \leq p} |X_{ij}|^3 \cdot 1 \left( \max_{1 \leq j \leq p} |X_{ij}| > \beta^{-1}/2 \right) \right] = C\beta^2\delta^{-1} B_3.$$

Therefore, we conclude that

$$|\mathbb{E}[f(S_n)] - \mathbb{E}[f(T_n)]| \leq C\beta\delta^{-1} \{B_1 + \beta(B_2 + B_3)\}.$$

**Step 4:** Combining Steps 1-3, one has

$$\begin{aligned} \mathbb{P}(Z \in A) &\leq (1 - \varepsilon)^{-1} \mathbb{E}[g \circ F_\beta(T_n)] + \frac{C\beta\delta^{-1} \{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon} \\ &\leq \mathbb{P}(F_\beta(T_n) \in A^{e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1} \{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon} \\ &\hspace{15em} \text{(by construction of } g) \\ &\leq \mathbb{P}(\tilde{Z}^* \in A^{2e_\beta + 3\delta}) + \frac{\varepsilon + C\beta\delta^{-1} \{B_1 + \beta(B_2 + B_3)\}}{1 - \varepsilon}. \quad \text{(by (17))} \end{aligned}$$

This completes the proof.  $\blacksquare$

**5. Inequalities for empirical processes.** In this section, we shall present some inequalities for empirical processes that will be used in the proofs of Theorem 2.1 and Lemma 2.2. These inequalities are of interest in their own rights. Consider the same setup as in Section 2, that is, let  $X_1, \dots, X_n$  be i.i.d. random variables taking values in a measurable space  $(S, \mathcal{S})$  with common distribution  $P$ . Let  $\mathcal{F}$  be a pointwise measurable class of functions  $S \rightarrow \mathbb{R}$ , to which a measurable envelope  $F$  is attached. In this section, however, we do not assume that  $\mathcal{F}$  is  $P$ -centered. Consider the empirical process  $\mathbb{G}_n f = n^{-1/2} \sum_{i=1}^n (f(X_i) - Pf)$ . Let  $\sigma^2 > 0$  be any positive constant such that  $\sup_{f \in \mathcal{F}} Pf^2 \leq \sigma^2 \leq \|F\|_{P,2}^2$ . Let  $M = \max_{1 \leq i \leq n} F(X_i)$ .

**THEOREM 5.1** (A useful deviation inequality for suprema of empirical processes). *Suppose that  $F \in \mathcal{L}^q(P)$  for some  $q \geq 2$ . Then for every  $t \geq 1$ , with probability  $> 1 - t^{-q/2}$ ,*

$$\begin{aligned} \|\mathbb{G}_n\|_{\mathcal{F}} &\leq (1 + \alpha) \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + K(q) \left[ (\sigma + n^{-1/2} \|M\|_q) \sqrt{t} \right. \\ &\quad \left. + \alpha^{-1} n^{-1/2} \|M\|_{2t} \right], \quad \forall \alpha > 0, \end{aligned}$$

where  $K(q) > 0$  is a constant depending only on  $q$ .

PROOF OF THEOREM 5.1. The theorem essentially follows from [4], Theorem 12, which states that

$$\|(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+\|_q \lesssim \sqrt{q}(\Sigma + \sigma) + qn^{-1/2}(\|M\|_q + \sigma),$$

where  $\Sigma^2 = \mathbb{E}[\|n^{-1} \sum_{i=1}^n (f(X_i) - Pf)^2\|_{\mathcal{F}}]$ . By Lemma 7 of the same paper,

$$\Sigma^2 \leq \sigma^2 + 64n^{-1/2}\|M\|_2\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + 32n^{-1}\|M\|_2^2.$$

Hence, using the simple inequality  $2\sqrt{ab} \leq \beta a + \beta^{-1}b, \forall \beta > 0$ , one has

$$\begin{aligned} \|(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+\|_q &\lesssim \sqrt{q}\beta\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + \sqrt{q}(1 + \beta^{-1})n^{-1/2}\|M\|_2 \\ &\quad + \sqrt{q}\sigma + qn^{-1/2}(\|M\|_q + \sigma). \end{aligned}$$

Therefore, by Markov's inequality, for every  $t \geq 1$ , with probability  $> 1 - t^{-q}$ ,

$$\begin{aligned} \|\mathbb{G}_n\|_{\mathcal{F}} &\leq \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + (\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}])_+ \\ &\leq (1 + C\sqrt{q}\beta t)\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] + C\sqrt{q}(1 + \beta^{-1})n^{-1/2}\|M\|_2 t \\ &\quad + C\sqrt{q}\sigma t + Cqn^{-1/2}(\|M\|_q + \sigma)t, \quad \forall \beta > 0. \end{aligned}$$

The final conclusion follows from taking  $\beta = C^{-1}q^{-1/2}t^{-1}\alpha$ .  $\blacksquare$

The proof of Lemma 2.2 relies on the following moment inequality for suprema of empirical processes, which is an extension of [60], Theorem 2.1, to possibly unbounded classes of functions (Theorem 3.1 of [60] derives a moment inequality applicable to the case where the envelope  $F$  has  $q > 4$  moments, but the form of the inequality in Theorem 5.2 is more convenient in our applications; note that Theorem 5.2 only requires  $F \in \mathcal{L}^2(P)$ , as opposed to  $F \in \mathcal{L}^q(P)$  with  $q > 4$  in Theorem 3.1 of [60], and Theorem 5.2 is not covered by [60]). Recall the uniform entropy integral  $J(\delta, \mathcal{F}, F)$ .

THEOREM 5.2 (A useful maximal inequality). *Suppose that  $F \in \mathcal{L}^2(P)$ . Let  $\delta = \sigma/\|F\|_{P,2}$ . Then*

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim J(\delta, \mathcal{F}, F)\|F\|_{P,2} + \frac{\|M\|_2 J^2(\delta, \mathcal{F}, F)}{\delta^2 \sqrt{n}}.$$

In the Supplemental Material [13], we give a full proof of Theorem 5.2 for the sake of completeness, although the proof is essentially similar to the proof of Theorem 2.1 in [60].

The bound in Theorem 5.2 will be explicit as soon as a suitable bound on the covering number is available. For example, the following corollary is an extension of [25], Proposition 2.1. For its proof, see Appendix A.5.

**COROLLARY 5.1** (Maximal inequality specialized to VC type classes). *Consider the same setup as in Theorem 5.2. Suppose that there exist constants  $A \geq e$  and  $v \geq 1$  such that  $\sup_Q N(\mathcal{F}, e_Q, \varepsilon \|F\|_{Q,2}) \leq (A/\varepsilon)^v$ ,  $0 < \forall \varepsilon \leq 1$ . Then*

$$\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}}] \lesssim \sqrt{v\sigma^2 \log\left(\frac{A\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_2}{\sqrt{n}} \log\left(\frac{A\|F\|_{P,2}}{\sigma}\right).$$

**6. Proof of Theorem 2.1.** We make use of Lemma 4.1 to prove the theorem. Construct a tight Gaussian random variable  $G_P$  in  $\ell^\infty(\mathcal{F})$  given in assumption (A3), independent of  $X_1, \dots, X_n$ . We note that one can extend  $G_P$  to the linear hull of  $\mathcal{F}$  in such a way that  $G_P$  has linear sample paths [see 19, Theorem 3.1.1]. Let  $\{f_1, \dots, f_N\}$  be a minimal  $\varepsilon\|F\|_{P,2}$ -net of  $(\mathcal{F}, e_P)$  with  $N = N(\mathcal{F}, e_P, \varepsilon\|F\|_{P,2})$ . Then for every  $f \in \mathcal{F}$ , there exists a function  $f_j, 1 \leq j \leq N$  such that  $e_P(f, f_j) < \varepsilon\|F\|_{P,2}$ . Recall  $\mathcal{F}_\varepsilon = \{f - g : f, g \in \mathcal{F}, e_P(f, g) < \varepsilon\|F\|_{P,2}\}$  and define

$$Z^\varepsilon = \max_{1 \leq j \leq N} \mathbb{G}_n f_j, \quad \tilde{Z}^* = \sup_{f \in \mathcal{F}} G_P f, \quad \tilde{Z}^{*\varepsilon} = \max_{1 \leq j \leq N} G_P f_j.$$

Observe that  $|Z - Z^\varepsilon| \leq \|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon}$  and  $|\tilde{Z}^{*\varepsilon} - \tilde{Z}^*| \leq \|G_P\|_{\mathcal{F}_\varepsilon}$ .

We shall apply Corollary 4.1 to  $Z^\varepsilon$ . Recall that  $\log(N \vee n) = H_n(\varepsilon)$ . Then for every Borel subset  $A$  of  $\mathbb{R}$  and  $\delta > 0$ ,

$$\mathbb{P}(Z^\varepsilon \in A) - \mathbb{P}(\tilde{Z}^{*\varepsilon} \in A^{16\delta}) \lesssim \delta^{-2} \{B_1 + \delta^{-1}(B_2 + B_4)H_n(\varepsilon)\} H_n(\varepsilon) + n^{-1} \log n,$$

where

$$\begin{aligned} B_1 &= n^{-1} \mathbb{E} \left[ \max_{1 \leq j, k \leq N} \left| \sum_{i=1}^n (f_j(X_i) f_k(X_i) - P(f_j f_k)) \right| \right], \\ B_2 &= n^{-3/2} \mathbb{E} \left[ \max_{1 \leq j \leq N} \sum_{i=1}^n |f_j(X_i)|^3 \right], \\ B_4 &= n^{-1/2} \mathbb{E} \left[ \max_{1 \leq j \leq N} |f_j(X_1)|^3 \cdot \mathbf{1} \left( \max_{1 \leq j \leq N} |f_j(X_1)| > \delta \sqrt{n} H_n(\varepsilon)^{-1} \right) \right]. \end{aligned}$$

Clearly  $B_1 \leq n^{-1/2} \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}, \mathcal{F}}]$ ,  $B_2 \leq n^{-1/2} \kappa^3$ , and  $B_4 \leq n^{-1/2} P[F^3 \mathbf{1}(F > \delta \sqrt{n} H_n(\varepsilon)^{-1})]$ . Hence choosing  $\delta > 0$  in such a way that

$$C\delta^{-2} n^{-1/2} \mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}, \mathcal{F}}] H_n(\varepsilon) \leq \frac{\gamma}{4}, \quad C\delta^{-3} n^{-1/2} \kappa^3 H_n^2(\varepsilon) \leq \frac{\gamma}{4},$$

that is,

$$\delta \geq C \max \left\{ \gamma^{-1/2} n^{-1/4} (\mathbb{E}[\|\mathbb{G}_n\|_{\mathcal{F}, \mathcal{F}}])^{1/2} H_n^{1/2}(\varepsilon), \gamma^{-1/3} n^{-1/6} \kappa H_n^{2/3}(\varepsilon) \right\},$$

we have

$$\mathbb{P}(Z^\varepsilon \in A) \leq \mathbb{P}(\tilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \frac{\gamma}{4}\kappa^{-3}P[F^3\mathbf{1}(F > \delta\sqrt{n}H_n(\varepsilon)^{-1})] + \frac{C \log n}{n}.$$

Note that  $\delta \geq c\gamma^{-1/3}n^{-1/6}\kappa H_n^{2/3}(\varepsilon)$ , so that

$$P[F^3\mathbf{1}(F > \delta\sqrt{n}H_n(\varepsilon)^{-1})] \leq P[F^3\mathbf{1}(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})].$$

Hence

$$\begin{aligned} \mathbb{P}(Z^\varepsilon \in A) &\leq \mathbb{P}(\tilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} \\ &\quad + \frac{\gamma}{4}P[(F/\kappa)^3\mathbf{1}(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\varepsilon)^{-1/3})] + \frac{C \log n}{n} \\ (26) \quad &=: \mathbb{P}(\tilde{Z}^{*\varepsilon} \in A^{16\delta}) + \frac{\gamma}{2} + \text{error}. \end{aligned}$$

By Theorem 5.1, with probability  $> 1 - \gamma/4$ ,

$$(27) \quad \|\mathbb{G}_n\|_{\mathcal{F}_\varepsilon} \leq K(q)\{\phi_n(\varepsilon) + (\varepsilon\|F\|_{P,2} + n^{-1/2}\|M\|_q)\gamma^{-1/q} + n^{-1/2}\|M\|_2\gamma^{-2/q}\} =: a,$$

where  $K(q)$  is a constant that depends only on  $q$ . Moreover, by the Borell-Sudakov-Tsirel'son inequality [59, Proposition A.1], with probability  $> 1 - \gamma/4$ , we have

$$(28) \quad \|G_P\|_{\mathcal{F}_\varepsilon} \leq \phi_n(\varepsilon) + \varepsilon\|F\|_{P,2}\sqrt{2\log(4/\gamma)} =: b.$$

Therefore, for every Borel subset  $A$  of  $\mathbb{R}$ ,

$$\begin{aligned} \mathbb{P}(Z \in A) &\leq \mathbb{P}(Z^\varepsilon \in A^a) + \frac{\gamma}{4} \quad (\text{by (27)}) \\ &\leq \mathbb{P}(\tilde{Z}^{*\varepsilon} \in A^{a+16\delta}) + \frac{3}{4}\gamma + \text{error} \quad (\text{by (26)}) \\ &\leq \mathbb{P}(\tilde{Z}^* \in A^{a+b+16\delta}) + \gamma + \text{error}. \quad (\text{by (28)}) \end{aligned}$$

The conclusion follows from Lemma 4.1. ■

**Acknowledgments.** The authors would like to thank the editors and anonymous referees for their careful review that helped improve upon the quality of the paper.

#### SUPPLEMENTARY MATERIAL

##### Supplement to ‘‘Gaussian approximation of suprema of empirical processes’’

( ). This supplemental file contains the additional technical proofs omitted in the main text, and some technical tools used in the proofs.

**References.**

- [1] Berthet, P. and Mason, D.M. (2006). Revisiting two strong approximation results of Dudley and Philipp. In: *High Dimensional Probability*, IMS Lecture Notes-Monograph Series, Vol. 51, pp.155-172.
- [2] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.
- [3] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- [4] Boucheron, S., Bousquet, O., Lugosi, G. and Massart, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514-560.
- [5] Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.* **17** 239-256.
- [6] Chatterjee, S. (2005). An error bound in the Sudakov-Fernique inequality. arXiv:math/0510424.
- [7] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257-283.
- [8] Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., and Wasserman, L. (2013). Stochastic convergence of persistence landscapes and silhouettes. arXiv:1312.0308.
- [9] Chen, L., Goldstein, L. and Shao, Q.-M. (2011). *Normal Approximation by Stein's Method*. Springer.
- [10] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807v3. To appear in *Probab. Theory and Related Fields*.
- [11] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786-2819.
- [12] Chernozhukov, V., Chetverikov, D. and Kato, K. (2013b). Anti-concentration and honest, adaptive confidence bands. arXiv:1303:7152.
- [13] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Supplement to "Gaussian approximation of suprema of empirical processes".
- [14] Chernozhukov, V., Lee, S., and Rosen, A. (2013). Intersection bounds: estimation and inference. *Econometrica* **81** 667-737.
- [15] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54-81.
- [16] Csörgo, M. and Horváth, L. (1993). *Weighted Approximations in Probability and Statistics*. Wiley.
- [17] Deheuvels, P. and Mason, D.M. (1994). Functional laws of the iterated logarithm for local empirical processes indexed by sets. *Ann. Probab.* **22** 1619-1661.
- [18] Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Z. Warhsch. Verw. Gabiete* **63** 393-432.
- [19] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.
- [20] Dudley, R.M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Warhsch. Verw. Gabiete* **62** 509-552.
- [21] Einmahl, U. and Mason, D.M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Theory Related Fields* **107** 283-311.
- [22] Einmahl, U. and Mason, D.M. (1998). Strong approximations to the local empirical process. In: *High Dimensional Probability* (eds. E. Eberlein, M. Hahn and M. Talagrand) pp. 75-92.

- [23] Einmahl, U. and Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1-37.
- [24] Einmahl, U. and Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380-1403.
- [25] Giné, E. and Guillou, A. (2001). On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. Inst. H. Poincaré Probab. Statist.* **37** 503-522.
- [26] Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **38** 907-921.
- [27] Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.* **37** 1605-1646.
- [28] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.
- [29] Ghosal, S., Sen, A. and van der Vaart, A.W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28** 1054-1082.
- [30] He, X. and Shao, Q.-M. (2000). On parameters on increasing dimensions. *J. Multivariate Anal.* **73** 125-135.
- [31] Huang, J.Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26** 242-272
- [32] Huang, J.Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statist. Probab. Lett.* **65** 207-216.
- [33] Kerkycharian, G., Nickl, R., and Picard, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probab. Theory Related Fields* **153** 363-404.
- [34] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Atti. Giorn.* **4** 83-91.
- [35] Koltchinskii, V.I. (1994). Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7** 73-118.
- [36] Komlós, J., Major, P., and Tusnády, G. (1975). An approximation for partial sums of independent rv's and the sample df I. *Z. Warhsch. Verw. Gabiete* **32** 111-131.
- [37] Konakov, V.D. and Piterbarg, V.I. (1984). On the convergence rate of maximal deviations distributions for kernel regression estimates. *J. Multivariate Anal.* **15** 279-294.
- [38] Le Cam, L. (1988). On the Prokhorov distance between the empirical process and the associated Gaussian bridge. Technical Report No. 170, Department of Statistics, University of California, Berkeley.
- [39] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces*. Springer.
- [40] Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.* **39** 201-231.
- [41] Mason, D.M. (2004). A uniform functional law of the logarithm for the local empirical process. *Ann. Probab.* **32** 1391-1418.
- [42] Mason, D.M. and van Zwet, W.R. (1987). A refinement of the KMT inequality for the uniform empirical process. *Ann. Probab.* **15** 871-884.
- [43] Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT construction. *Ann. Probab.* **17** 266-291.
- [44] Meckes, E. (2009). On Stein's method for multivariate normal approximation. In: *High Dimensional Probability V: The Luminy Volume*, IMS Collections, Vol.5, pp.159-178.
- [45] Neumann, M. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* **26** 2014-2048.

- [46] Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147-168.
- [47] Nolan, D. and Pollard, D. (1987).  $U$ -processes: rates of convergence. *Ann. Statist.* **15** 780-799.
- [48] Norvaiša, R. and Paulauskas, V. (1991). Rate of convergence in the Central Limit Theorem for empirical processes. *J.Theoret. Probab.* **4** 511-534.
- [49] Panchenko, D. (2013). *The Sherrington-Kirkpatrick Model*. Springer-Vegrlag, New York.
- [50] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.
- [51] Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37** 2150-2173.
- [52] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.
- [53] Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.* **41** 1299-1328.
- [54] Settati, A. (2009). Gaussian approximation of the empirical process under random entropy conditions. *Stochastic Process. Appl.* **119** 1541-1560.
- [55] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: *Proc. of the Sixth Berkeley Symp. on Math. Statist. and Probab.*, Vol. II: Probability theory. pp.583-602.
- [56] Stein, C. (1986). *Approximate Computation of Expectations*. IMS Lecture Notes-Monograph Series, Vol.7.
- [57] Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505-563.
- [58] Talagrand, M. (2005). *The Generic Chaining*. Springer.
- [59] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- [60] van der Vaart, A.W. and Wellner, J.A. (2011). A local maximal inequality under uniform entropy. *Electronic J. Statist.* **5** 192-203.
- [61] Yurinskii, V.V. (1977). On the error of the Gaussian approximation for convolutions. *Theory of Probability and Its Applications* **2** 236-247.
- [62] Zaitsev, Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions. *Probab. Theory Related Fields* **74** 535-566.

DEPARTMENT OF ECONOMICS AND  
OPERATIONS RESEARCH CENTER, MIT  
50 MEMORIAL DRIVE  
CAMBRIDGE, MA 02142, USA.  
E-MAIL: vchern@mit.edu

DEPARTMENT OF ECONOMICS, UCLA  
BUNCHE HALL, 8283  
315 PORTOLA PLAZA  
LOS ANGELES, CA 90095, USA.  
E-MAIL: chetverikov@econ.ucla.edu

GRADUATE SCHOOL OF ECONOMICS  
UNIVERSITY OF TOKYO  
7-3-1 HONGO, BUNKYO-KU  
TOKYO 113-0033, JAPAN.  
E-MAIL: kkato@e.u-tokyo.ac.jp